



To our friends and colleagues

Hiroyuki Shima · Tsuneyoshi Nakayama

# Higher Mathematics for Physics and Engineering

 Springer

---

## Preface

Owing to the rapid advances in the physical sciences and engineering, the demand for higher-level mathematics is increasing yearly. This book is designed for advanced undergraduates and graduate students who are interested in the mathematical aspects of their own fields of study. The reader is assumed to have a knowledge of undergraduate-level calculus and linear algebra.

There are any number of books available on mathematics for physics and engineering but they all fall into one of two categories: the one emphasizes mathematical rigor and the exposition of definitions or theorems, whereas the other is concerned primarily with applying mathematics to practical problems. We believe that neither of these approaches alone is particularly helpful to physicists and engineers who want to understand the mathematical background of the subjects with which they are concerned. This book is different in that it provides a short path to higher mathematics via a combination of these approaches. A sizable portion of this book is devoted to theorems and definitions with their proofs, and we are convinced that the study of these proofs, which range from trivial to difficult, is useful for a grasp of the general idea of mathematical logic. Moreover, several problems have been included at the end of each section, and complete solutions for all of them are presented in the greatest possible detail. We firmly believe that ours is a better pedagogical approach than that found in typical textbooks, where there are many well-polished problems but no solutions.

This book is essentially self-contained and assumes only standard undergraduate preparation such as elementary calculus and linear algebra. The first half of the book covers the following three topics: real analysis, functional analysis, and complex analysis, along with the preliminaries and four appendixes. Part I focuses on sequences and series of real numbers of real functions, with detailed explanations of their convergence properties. We also emphasize the concepts of Cauchy sequences and the Cauchy criterion that determine the convergence of infinite real sequences. Part II deals with the theory of the Hilbert space, which is the most important class of infinite vector spaces. The completeness property of Hilbert spaces allows one to develop

various types of complex orthonormal polynomials, as described in the middle of Part II. An introduction to the Lebesgue integration theory, a subject of ever-increasing importance in physics, is also presented. Part III describes the theory of complex-valued functions of one complex variable. All relevant elements including analytic functions, singularity, residue, continuation, and conformal mapping are described in a self-contained manner. A thorough understanding of the fundamentals treated is important in order to proceed to more advanced branches of mathematical physics.

In the second half of the volume, the following three specific topics are discussed: Fourier analysis, differential equations, and tensor analysis. These three are the most important subjects in both engineering and the physical sciences, but their rigorous mathematical structures have hardly been covered in ordinary textbooks. We know that mathematical rigor is often unnecessary for practical use. However, the blind usage of mathematical methods as a tool may lead to a lack of understanding of the symbiotic relationship between mathematics and the physical sciences. We believe that readers who study the mathematical structures underlying these three subjects in detail will acquire a better understanding of the theoretical backgrounds associated with their own fields. Part IV describes the theory of Fourier series, the Fourier transform, and the Laplace transform, with a special emphasis on the proofs of their convergence properties. A more contemporary subject, the wavelet transform, is also described toward the end of Part IV. Part V deals with ordinary and partial differential equations. The existence theorem and stability theory for solutions, which serve as the underlying basis for differential equations, are described with rigorous proofs. Part VI is devoted to the calculus of tensors in terms of both Cartesian and non-Cartesian coordinates, along with the essentials of differential geometry. An alternative tensor theory expressed in terms of abstract vector spaces is developed toward the end of Part VI.

The authors hope and trust that this book will serve as an introductory guide for the mathematical aspects of the important topics in the physical sciences and engineering.

Sapporo,  
November 2009

*Hiroyuki Shima*  
*Tsuneyoshi Nakayama*

---

# Contents

<b>1</b>	<b>Preliminaries</b>	<b>1</b>
1.1	Basic Notions of a Set	1
1.1.1	Set and Element	1
1.1.2	Number Sets	3
1.1.3	Bounds	3
1.1.4	Interval	4
1.1.5	Neighborhood and Contact Point	5
1.1.6	Closed and Open Sets	7
1.2	Conditional Statements	9
1.3	Order of Magnitude	10
1.3.1	Symbols $O$ , $o$ , and $\sim$	10
1.3.2	Asymptotic Behavior	11
1.4	Values of Indeterminate Forms	12
1.4.1	l'Hôpital's Rule	12
1.4.2	Several Examples	13

---

## Part I Real Analysis

---

<b>2</b>	<b>Real Sequences and Series</b>	<b>17</b>
2.1	Sequences of Real Numbers	17
2.1.1	Convergence of a Sequence	17
2.1.2	Bounded Sequences	18
2.1.3	Monotonic Sequences	19
2.1.4	Limit Superior and Limit Inferior	21
2.2	Cauchy Criterion for Real Sequences	25
2.2.1	Cauchy Sequence	25
2.2.2	Cauchy Criterion	26
2.3	Infinite Series of Real Numbers	29
2.3.1	Limits of Infinite Series	29
2.3.2	Cauchy Criterion for Infinite Series	31

2.3.3	Absolute and Conditional Convergence . . . . .	32
2.3.4	Rearrangements . . . . .	34
2.4	Convergence Tests for Infinite Real Series . . . . .	38
2.4.1	Limit Tests . . . . .	38
2.4.2	Ratio Tests . . . . .	40
2.4.3	Root Tests . . . . .	41
2.4.4	Alternating Series Test . . . . .	42
<b>3</b>	<b>Real Functions</b> . . . . .	<b>45</b>
3.1	Fundamental Properties . . . . .	45
3.1.1	Limit of a Function . . . . .	45
3.1.2	Continuity of a Function . . . . .	47
3.1.3	Derivative of a Function . . . . .	48
3.1.4	Smooth Functions . . . . .	50
3.2	Sequences of Real Functions . . . . .	50
3.2.1	Pointwise Convergence . . . . .	50
3.2.2	Uniform Convergence . . . . .	52
3.2.3	Cauchy Criterion for Series of Functions . . . . .	53
3.2.4	Continuity of the Limit Function . . . . .	54
3.2.5	Integrability of the Limit Function . . . . .	56
3.2.6	Differentiability of the Limit Function . . . . .	57
3.3	Series of Real Functions . . . . .	61
3.3.1	Series of Functions . . . . .	61
3.3.2	Properties of Uniformly Convergent Series of Functions . . . . .	62
3.3.3	Weierstrass $M$ -test . . . . .	63
3.4	Improper Integrals . . . . .	66
3.4.1	Definitions . . . . .	66
3.4.2	Convergence of an Improper Integral . . . . .	67
3.4.3	Principal Value Integral . . . . .	67
3.4.4	Conditions for Convergence . . . . .	68

---

## Part II Functional Analysis

---

<b>4</b>	<b>Hilbert Spaces</b> . . . . .	<b>73</b>
4.1	Hilbert Spaces . . . . .	73
4.1.1	Introduction . . . . .	73
4.1.2	Abstract Vector Spaces . . . . .	74
4.1.3	Inner Product . . . . .	75
4.1.4	Geometry of Inner Product Spaces . . . . .	76
4.1.5	Orthogonality . . . . .	78
4.1.6	Completeness of Vector Spaces . . . . .	79
4.1.7	Several Examples of Hilbert Spaces . . . . .	80
4.2	Hierarchical Structure of Vector Spaces . . . . .	83
4.2.1	Precise Definitions of Vector Spaces . . . . .	83

4.2.2	Metric Space . . . . .	84
4.2.3	Normed Spaces . . . . .	85
4.2.4	Subspaces of a Normed Space . . . . .	86
4.2.5	Basis of a Vector Space: Revisited . . . . .	87
4.2.6	Orthogonal Bases in Hilbert Spaces . . . . .	88
4.3	Hilbert Spaces of $\ell^2$ and $L^2$ . . . . .	91
4.3.1	Completeness of the $\ell^2$ Spaces . . . . .	91
4.3.2	Completeness of the $L^2$ Spaces . . . . .	92
4.3.3	Mean Convergence . . . . .	95
4.3.4	Generalized Fourier Coefficients . . . . .	95
4.3.5	Riesz–Fisher Theorem . . . . .	96
4.3.6	Isomorphism between $\ell^2$ and $L^2$ . . . . .	98
<b>5</b>	<b>Orthonormal Polynomials</b> . . . . .	<b>101</b>
5.1	Polynomial Approximations . . . . .	101
5.1.1	Weierstrass Theorem . . . . .	101
5.1.2	Existence of Complete Orthonormal sets of Polynomials . . . . .	103
5.1.3	Legendre Polynomials . . . . .	105
5.1.4	Fourier Series . . . . .	108
5.1.5	Spherical Harmonic Functions . . . . .	109
5.2	Classification of Orthonormal Functions . . . . .	114
5.2.1	General Rodrigues Formula . . . . .	114
5.2.2	Classification of the Polynomials . . . . .	116
5.2.3	The Recurrence Formula . . . . .	119
5.2.4	Coefficients of the Recurrence Formula . . . . .	120
5.2.5	Roots of Orthogonal Polynomials . . . . .	121
5.2.6	Differential Equations Satisfied by the Polynomials . . . . .	122
5.2.7	Generating Functions (I) . . . . .	124
5.2.8	Generating Functions (II) . . . . .	125
5.3	Chebyshev Polynomials . . . . .	128
5.3.1	Minimax Property . . . . .	128
5.3.2	A Concise Representation . . . . .	131
5.3.3	Discrete Orthogonality Relation . . . . .	133
5.4	Applications in Physics and Engineering . . . . .	135
5.4.1	Quantum-Mechanical State in an Harmonic Potential . . . . .	135
5.4.2	Electrostatic potential generated by a multipole . . . . .	136
<b>6</b>	<b>Lebesgue Integrals</b> . . . . .	<b>139</b>
6.1	Measure and Summability . . . . .	139
6.1.1	Riemann Integral Revisited . . . . .	139
6.1.2	Measure . . . . .	141
6.1.3	The Probability Measure . . . . .	142
6.1.4	Support and Area of a Step Function . . . . .	144
6.1.5	$\alpha$ -Summability . . . . .	146
6.1.6	Properties of $\alpha$ -summable functions . . . . .	147

6.2	Lebesgue Integral . . . . .	149
6.2.1	Lebesgue Measure . . . . .	149
6.2.2	Definition of the Lebesgue Integral . . . . .	151
6.2.3	Riemann Integrals vs. Lebesgue Integrals . . . . .	152
6.2.4	Properties of the Lebesgue Integrals . . . . .	153
6.2.5	Null-Measure Property of Countable Sets . . . . .	154
6.2.6	The Concept of Almost Everywhere . . . . .	155
6.3	Important Theorems for Lebesgue Integrals . . . . .	158
6.3.1	Monotone Convergence Theorem . . . . .	158
6.3.2	Dominated Convergence Theorem (I) . . . . .	160
6.3.3	Fatou Lemma . . . . .	160
6.3.4	Dominated Convergence Theorem (II) . . . . .	161
6.3.5	Fubini Theorem . . . . .	162
6.4	The Lebesgue Spaces $L^p$ . . . . .	167
6.4.1	The Spaces of $L^p$ . . . . .	167
6.4.2	Hölder Inequality . . . . .	168
6.4.3	Minkowski Inequality . . . . .	169
6.4.4	Completeness of $L^p$ Spaces . . . . .	170
6.5	Applications in Physics and Engineering . . . . .	172
6.5.1	Practical Significance of Lebesgue Integrals . . . . .	172
6.5.2	Contraction Mapping . . . . .	173
6.5.3	Preliminaries for the Central Limit Theorem . . . . .	175
6.5.4	Central Limit Theorem . . . . .	177
6.5.5	Proof of the Central Limit Theorem . . . . .	178

---

## Part III Complex Analysis

---

7	<b>Complex Functions</b> . . . . .	185
7.1	Analytic Functions . . . . .	185
7.1.1	Continuity and Differentiability . . . . .	185
7.1.2	Definition of an Analytic Function . . . . .	187
7.1.3	Cauchy–Riemann Equations . . . . .	189
7.1.4	Harmonic Functions . . . . .	191
7.1.5	Geometric Interpretation of Analyticity . . . . .	192
7.2	Complex Integrations . . . . .	195
7.2.1	Integration of Complex Functions . . . . .	195
7.2.2	Cauchy Theorem . . . . .	197
7.2.3	Integrations on a Multiply Connected Region . . . . .	199
7.2.4	Primitive Functions . . . . .	201
7.3	Cauchy Integral Formula and Related Theorem . . . . .	204
7.3.1	Cauchy Integral Formula . . . . .	204
7.3.2	Goursat Formula . . . . .	206
7.3.3	Absence of Extrema in Analytic Regions . . . . .	207
7.3.4	Liouville Theorem . . . . .	208

7.3.5	Fundamental Theorem of Algebra .....	209
7.3.6	Morera Theorem .....	210
7.4	Series Representations .....	213
7.4.1	Circle of Convergence .....	213
7.4.2	Singularity on the Radius of Convergence.....	215
7.4.3	Taylor Series.....	217
7.4.4	Apparent Paradoxes .....	218
7.4.5	Laurent Series .....	219
7.4.6	Regular and Principal Parts .....	221
7.4.7	Uniqueness of Laurent Series.....	222
7.4.8	Techniques for Laurent Expansion .....	223
7.5	Applications in Physics and Engineering.....	228
7.5.1	Fluid Dynamics .....	228
7.5.2	Kutta–Joukowski Theorem .....	229
7.5.3	Blasius Formula .....	231
<b>8</b>	<b>Singularity and Continuation .....</b>	<b>233</b>
8.1	Singularity .....	233
8.1.1	Isolated Singularities.....	233
8.1.2	Nonisolated Singularities .....	235
8.1.3	Weierstrass Theorem for Essential Singularities.....	236
8.1.4	Rational Functions .....	237
8.2	Multivaluedness .....	240
8.2.1	Multivalued Functions .....	240
8.2.2	Riemann Surfaces .....	241
8.2.3	Branch Point and Branch Cut .....	243
8.3	Analytic Continuation .....	245
8.3.1	Continuation by Taylor Series.....	245
8.3.2	Function Elements.....	246
8.3.3	Uniqueness Theorem.....	250
8.3.4	Conservation of Functional Equations .....	250
8.3.5	Continuation Around a Branch Point .....	252
8.3.6	Natural Boundaries .....	252
8.3.7	Technique of Analytic Continuations .....	254
8.3.8	The Method of Moment .....	255
<b>9</b>	<b>Contour Integrals .....</b>	<b>259</b>
9.1	Calculus of Residues .....	259
9.1.1	Residue Theorem.....	259
9.1.2	Remarks on Residues .....	261
9.1.3	Winding Number.....	262
9.1.4	Ratio Method.....	263
9.1.5	Evaluating the Residues .....	264
9.2	Applications to Real Integrals.....	267
9.2.1	Classification of Evaluable Real Integrals .....	267

9.2.2	Type 1: Integrals of $f(\cos \theta, \sin \theta)$ . . . . .	268
9.2.3	Type 2: Integrals of Rational Function . . . . .	268
9.2.4	Type 3: Integrals of $f(x)e^{ix}$ . . . . .	270
9.2.5	Type 4: Integrals of $f(x)/x^\alpha$ . . . . .	271
9.2.6	Type 5: Integrals of $f(x) \log x$ . . . . .	273
9.3	More Applications of Residue Calculus . . . . .	277
9.3.1	Integrals on Rectangular Contours . . . . .	277
9.3.2	Fresnel Integrals . . . . .	279
9.3.3	Summation of Series . . . . .	281
9.3.4	Langevin and Riemann zeta Functions . . . . .	283
9.4	Argument Principle . . . . .	285
9.4.1	The Principle . . . . .	285
9.4.2	Variation of the Argument . . . . .	288
9.4.3	Extentson of the Argument Principle . . . . .	289
9.4.4	Rouché Theorem . . . . .	290
9.5	Dispersion Relations . . . . .	293
9.5.1	Principal Value Integrals . . . . .	293
9.5.2	Several Remarks . . . . .	295
9.5.3	Dispersion relations . . . . .	297
9.5.4	Kramers–Kronig Relations . . . . .	298
9.5.5	Subtracted Dispersion Relation . . . . .	299
9.5.6	Derivation of Dispersion Relations . . . . .	300
<b>10</b>	<b>Conformal Mapping</b> . . . . .	<b>305</b>
10.1	Fundamentals . . . . .	305
10.1.1	Conformal Property of Analytic Functions . . . . .	305
10.1.2	Scale Factor . . . . .	307
10.1.3	Mapping of a Differential Area . . . . .	308
10.1.4	Mapping of a Tangent Line . . . . .	309
10.1.5	The Point at Infinity . . . . .	311
10.1.6	Singular Point at Infinity . . . . .	312
10.2	Elementary Transformations . . . . .	315
10.2.1	Linear Transformations . . . . .	315
10.2.2	Bilinear Transformations . . . . .	316
10.2.3	Miscellaneous Transformations . . . . .	317
10.2.4	Mapping of Finite-Radius Circle . . . . .	321
10.2.5	Invariance of the Cross ratio . . . . .	322
10.3	Applications to Boundary-Value Problems . . . . .	325
10.3.1	Schwarz–Christoffel Transformation . . . . .	325
10.3.2	Derivation of the Schwartz–Christoffel Transformation . . . . .	326
10.3.3	The Method of Inversion . . . . .	327
10.4	Applications in Physics and Engineering . . . . .	332
10.4.1	Electric Potential Field in a Complicated Geometry . . . . .	332
10.4.2	Joukowski Airfoil . . . . .	335

---

**Part IV Fourier Analysis**


---

<b>11</b>	<b>Fourier Series</b>	339
11.1	Basic Properties	339
11.1.1	Definition	339
11.1.2	Dirichlet Theorem	340
11.1.3	Fourier Series of Periodic Functions	342
11.1.4	Half-range Fourier Series	343
11.1.5	Fourier Series of Nonperiodic Functions	344
11.1.6	The Rate of Convergence	346
11.1.7	Fourier Series in Higher Dimensions	347
11.2	Mean Convergence of Fourier Series	351
11.2.1	Mean Convergence Property	351
11.2.2	Dirichlet and Fejér Integrals	353
11.2.3	Proof of the Mean Convergence of Fourier Series	355
11.2.4	Parseval Identity	356
11.2.5	Riemann–Lebesgue Theorem	357
11.3	Uniform Convergence of Fourier series	360
11.3.1	Criterion for Uniform and Pointwise Convergence	360
11.3.2	Fejér theorem	360
11.3.3	Proof of Uniform Convergence	361
11.3.4	Pointwise Convergence at Discontinuous Points	363
11.3.5	Gibbs Phenomenon	365
11.3.6	Overshoot at a Discontinuous Point	366
11.4	Applications in Physics and Engineering	371
11.4.1	Temperature Variation of the Ground	371
11.4.2	String Vibration Under Impact	373
<b>12</b>	<b>Fourier Transformation</b>	377
12.1	Fourier Transform	377
12.1.1	Derivation of Fourier Transform	377
12.1.2	Fourier Integral Theorem	379
12.1.3	Proof of the Fourier Integral Theorem	380
12.1.4	Inverse Relations of the Half-width	381
12.1.5	Parseval Identity for Fourier Transforms	382
12.1.6	Fourier Transforms in Higher Dimensions	384
12.2	Convolution and Correlations	387
12.2.1	Convolution Theorem	387
12.2.2	Cross-Correlation Functions	388
12.2.3	Autocorrelation Functions	390
12.3	Discrete Fourier Transform	391
12.3.1	Definitions	391
12.3.2	Inverse Transform	392
12.3.3	Nyquist Frequency and Aliasing	393

12.3.4	Sampling Theorem .....	394
12.3.5	Fast Fourier Transform.....	396
12.3.6	Matrix Representation of FFT Algorithm.....	398
12.3.7	Decomposition Method for FFT .....	400
12.4	Applications in Physics and Engineering.....	401
12.4.1	Fraunhofer Diffraction I .....	401
12.4.2	Fraunhofer Diffraction II .....	403
12.4.3	Amplitude Modulation Technique .....	404
<b>13</b>	<b>Laplace Transformation .....</b>	<b>407</b>
13.1	Basic Operations .....	407
13.1.1	Definitions .....	407
13.1.2	Several Remarks .....	408
13.1.3	Significance of Analytic Continuation .....	409
13.1.4	Convergence of Laplace Integrals .....	410
13.1.5	Abscissa of Absolute Convergence .....	411
13.1.6	Laplace Transforms of Elementary Functions.....	412
13.2	Properties of Laplace Transforms .....	415
13.2.1	First Shifting Theorem .....	415
13.2.2	Second Shifting Theorem .....	416
13.2.3	Laplace Transform of Periodic Functions .....	417
13.2.4	Laplace Transform of Derivatives and Integrals .....	418
13.2.5	Laplace Transforms Leading to Multivalued Functions ..	420
13.3	Convergence Theorems for Laplace Integrals .....	422
13.3.1	Functions of Exponential Order .....	422
13.3.2	Convergence for Exponential-Order Cases .....	424
13.3.3	Uniform Convergence for Exponential-Order Cases ....	425
13.3.4	Convergence for General Cases .....	427
13.3.5	Uniform Convergence for General Cases .....	429
13.3.6	Distinction Between Exponential-Order Cases and General Cases.....	431
13.3.7	Analytic Property of Laplace Transforms .....	432
13.4	Inverse Laplace Transform .....	432
13.4.1	The Two-Sided Laplace Transform .....	432
13.4.2	Inverse of the Two-Sided Laplace Transform .....	434
13.4.3	Inverse of the One-Sided Laplace Transform .....	436
13.4.4	Useful Formula for Inverse Laplace Transformation ....	436
13.4.5	Evaluating Inverse Transformations .....	439
13.4.6	Inverse Transform of Multivalued Functions.....	441
13.5	Applications in Physics and Engineering.....	445
13.5.1	Electric Circuits I .....	445
13.5.2	Electric Circuits II .....	447

<b>14 Wavelet Transformation</b>	449
14.1 Continuous Wavelet Analyses	449
14.1.1 Definition of Wavelet	449
14.1.2 The Wavelet Transform	451
14.1.3 Correlation Between Wavelet and Signal	452
14.1.4 Actual Application of the Wavelet Transform	455
14.1.5 Inverse Wavelet Transform	456
14.1.6 Noise Reduction Technique	457
14.2 Discrete Wavelet Analysis	460
14.2.1 Discrete Wavelet Transforms	460
14.2.2 Complete Orthonormal Wavelets	462
14.2.3 Multiresolution Analysis	463
14.2.4 Orthogonal Decomposition	464
14.2.5 Constructing an Orthonormal Basis	466
14.2.6 Two-Scale Relations	467
14.2.7 Constructing the Mother Wavelet	469
14.2.8 Multiresolution Representation	471
14.3 Fast Wavelet Transformation	476
14.3.1 Generalized Two-Scale Relations	476
14.3.2 Decomposition Algorithm	478
14.3.3 Reconstruction Algorithm	479

---

## Part V Differential Equations

---

<b>15 Ordinary Differential Equations</b>	483
15.1 Concepts of Solutions	483
15.1.1 Definition of Ordinary Differential Equations	483
15.1.2 Explicit Solution	484
15.1.3 Implicit Solution	485
15.1.4 General and Particular Solutions	486
15.1.5 Singular Solution	488
15.1.6 Integral Curve and Direction Field	489
15.2 Existence Theorem for the First-Order ODE	491
15.2.1 Picard Method	491
15.2.2 Properties of Successive Approximations	493
15.2.3 Existence Theorem and Lipschitz Condition	495
15.2.4 Uniqueness Theorem	497
15.2.5 Remarks on the Two Theorems	498
15.3 Sturm–Liouville Problems	500
15.3.1 Sturm–Liouville Equation	500
15.3.2 Conversion into a Sturm–Liouville Equation	501
15.3.3 Self-adjoint Operators	502
15.3.4 Required Boundary Condition	503
15.3.5 Reality of Eigenvalues	504

<b>16</b>	<b>System of Ordinary Differential Equations</b>	509
16.1	Systems of ODEs	509
16.1.1	Systems of the First-Order ODEs	509
16.1.2	Column-Vector Notation	510
16.1.3	Reducing the Order of ODEs	510
16.1.4	Lipschitz Condition in Vector Spaces	512
16.2	Linear System of ODEs	513
16.2.1	Basic Terminology	513
16.2.2	Vector Space of Solutions	514
16.2.3	Fundamental Systems of Solutions	516
16.2.4	Wronskian for a System of ODEs	517
16.2.5	Liouville Formula for a Wronskian	518
16.2.6	Wronskian for an $n$ th-Order Linear ODE	519
16.2.7	Particular Solution of an Inhomogeneous System	522
16.3	Autonomous Systems of ODEs	525
16.3.1	Autonomous System	525
16.3.2	Trajectory	526
16.3.3	Critical Point	527
16.3.4	Stability of a Critical Point	527
16.3.5	Linear Autonomous System	528
16.4	Classification of Critical Points	530
16.4.1	Improper Node	530
16.4.2	Saddle Point	531
16.4.3	Proper Node	532
16.4.4	Spiral Point	533
16.4.5	Center	533
16.4.6	Limit Cycle	534
16.5	Applications in Physics and Engineering	536
16.5.1	Van der Pol Generator	536
<b>17</b>	<b>Partial Differential Equations</b>	539
17.1	Basic Properties	539
17.1.1	Definitions	539
17.1.2	Subsidiary Conditions	540
17.1.3	Linear and Homogeneous PDEs	540
17.1.4	Characteristic Equation	541
17.1.5	Second-Order PDEs	543
17.1.6	Classification of Second-Order PDEs	544
17.2	The Laplacian Operator	546
17.2.1	Maximum and Minimum Theorem	546
17.2.2	Uniqueness Theorem	548
17.2.3	Symmetric Properties of the Laplacian	548
17.3	The Diffusion Operator	550
17.3.1	The Diffusion Equations in Bounded Domains	550
17.3.2	Maximum and Minimum Theorem	551
17.3.3	Uniqueness Theorem	551

17.4	The Wave Operator . . . . .	552
17.4.1	The Cauchy Problem . . . . .	552
17.4.2	Homogeneous Wave Equations . . . . .	553
17.4.3	Inhomogeneous Wave Equations . . . . .	555
17.4.4	Wave Equations in Finite Domains . . . . .	556
17.5	Applications in Physics and Engineering . . . . .	559
17.5.1	Wave Equations for Vibrating Strings . . . . .	559
17.5.2	Diffusion Equations for Heat Conduction . . . . .	561

---

## Part VI Tensor Analyses

---

<b>18</b>	<b>Cartesian Tensors . . . . .</b>	<b>565</b>
18.1	Rotation of Coordinate Axes . . . . .	565
18.1.1	Tensors and Coordinate Transformations . . . . .	565
18.1.2	Summation Convention . . . . .	566
18.1.3	Cartesian Coordinate System . . . . .	567
18.1.4	Rotation of Coordinate Axes . . . . .	568
18.1.5	Orthogonal Relations . . . . .	569
18.1.6	Matrix Representations . . . . .	570
18.1.7	Determinant of a Matrix . . . . .	571
18.2	Cartesian Tensors . . . . .	576
18.2.1	Cartesian Vectors . . . . .	576
18.2.2	A Vector and a Geometric Arrow . . . . .	577
18.2.3	Cartesian Tensors . . . . .	578
18.2.4	Scalars . . . . .	579
18.3	Pseudotensors . . . . .	580
18.3.1	Improper Rotations . . . . .	580
18.3.2	Pseudovectors . . . . .	582
18.3.3	Pseudotensors . . . . .	584
18.3.4	Levi-Civita Symbols . . . . .	584
18.4	Tensor Algebra . . . . .	586
18.4.1	Addition and Subtraction . . . . .	586
18.4.2	Contraction . . . . .	587
18.4.3	Outer and Inner Products . . . . .	587
18.4.4	Symmetric and Antisymmetric Tensors . . . . .	589
18.4.5	Equivalence of an Antisymmetric Second-Order Tensor to a Pseudovector . . . . .	590
18.4.6	Quotient Theorem . . . . .	592
18.4.7	Quotient Theorem for Two-Subscripted Quantities . . . . .	593
18.5	Applications in Physics and Engineering . . . . .	596
18.5.1	Inertia Tensor . . . . .	596
18.5.2	Tensors in Electromagnetism in Solids . . . . .	598
18.5.3	Electromagnetic Field Tensor . . . . .	598
18.5.4	Elastic Tensor . . . . .	600

<b>19</b>	<b>Non-Cartesian Tensors</b>	601
19.1	Curvilinear Coordinate Systems	601
19.1.1	Local Basis Vectors	601
19.1.2	Reciprocity Relations	603
19.1.3	Transformation Law of Covariant Basis Vectors	604
19.1.4	Transformation Law of Contravariant Basis Vectors	606
19.1.5	Components of a Vector	606
19.1.6	Components of a Tensor	608
19.1.7	Mixed Components of a Tensor	609
19.1.8	Kronecker Delta	610
19.2	Metric Tensor	611
19.2.1	Definition	611
19.2.2	Geometric Role of Metric Tensors	612
19.2.3	Riemann Space and Metric Tensor	613
19.2.4	Elements of Arc, Area, and Volume	614
19.2.5	Scale Factors	616
19.2.6	Representation of Basis Vectors in Derivatives	617
19.2.7	Index Lowering and Raising	617
19.3	Christoffel Symbols	621
19.3.1	Derivatives of Basis Vectors	621
19.3.2	Nontensor Character	622
19.3.3	Properties of Christoffel Symbols	623
19.3.4	Alternative Expression	623
19.4	Covariant Derivatives	627
19.4.1	Covariant Derivatives of Vectors	627
19.4.2	Remarks on Covariant Derivatives	628
19.4.3	Covariant Derivatives of Tensors	629
19.4.4	Vector Operators in Tensor Form	630
19.5	Applications in Physics and Engineering	634
19.5.1	General Relativity Theory	634
19.5.2	Riemann Tensor	635
19.5.3	Energy–Momentum Tensor	636
19.5.4	Einstein Field Equation	637
<b>20</b>	<b>Tensor as Mapping</b>	639
20.1	Vector as a Linear Function	639
20.1.1	Overview	639
20.1.2	Vector Spaces Revisited	640
20.1.3	Vector Spaces of Linear Functions	640
20.1.4	Dual Spaces	641
20.1.5	Equivalence Between Vectors and Linear Functions	642
20.2	Tensor as Multilinear Function	643
20.2.1	Direct Product of Vector Spaces	643
20.2.2	Multilinear Functions	644
20.2.3	Tensor Product	644

20.2.4	General Definition of Tensors .....	645
20.3	Components of Tensors .....	646
20.3.1	Basis of a Tensor Space .....	646
20.3.2	Transformation Laws of Tensors .....	648
20.3.3	Natural Isomorphism .....	648
20.3.4	Inner Product in Tensor Language .....	651
20.3.5	Index Lowering and Raising in Tensor Language .....	652

---

## Part VII Appendixes

---

<b>A</b>	<b>Proof of the Bolzano–Weierstrass Theorem .....</b>	<b>657</b>
A.1	Limit Points .....	657
A.2	Cantor Theorem .....	658
A.3	Bolzano–Weierstrass Theorem .....	659
<b>B</b>	<b>Dirac <math>\delta</math> Function .....</b>	<b>661</b>
B.1	Basic Properties .....	661
B.2	Representation as a Limit of Function .....	662
B.3	Remarks on Representation 4 .....	663
<b>C</b>	<b>Proof of Weierstrass Approximation Theorem .....</b>	<b>667</b>
<b>D</b>	<b>Tabulated List of Orthonormal Polynomial Functions .....</b>	<b>671</b>
	<b>Index .....</b>	<b>677</b>



## Preliminaries

This chapter provides the basic notation, terminology, and abbreviations that we will be using, particularly in real analysis, and is designed to serve as a reference rather than as a systematic exposition.

### 1.1 Basic Notions of a Set

#### 1.1.1 Set and Element

A **set** is a collection of **elements** (or **points**) that are definite and separate objects. If  $a$  is an element of a set  $S$ , we write

$$a \in S.$$

Otherwise, we write

$$a \notin S$$

to indicate that  $a$  does not belong to  $S$ . If a set contains no elements, it is called an **empty set** and is designated by  $\emptyset$ .

A set may be defined by listing its elements or by providing a rule that determines which elements belong to it. For example, we write

$$X = \{x_1, x_2, x_3, \dots, x_n\}$$

to indicate that  $X$  is a set with  $n$  elements:  $x_1, x_2, \dots, x_n$ . When a set contains a finite (infinite) number of elements, it is called a **finite** (**infinite set**).

A set  $X$  is said to be a **subset** of  $Y$  if every element in  $X$  is also an element in  $Y$ . This relationship is expressed as

$$X \subseteq Y.$$

When  $X \subseteq Y$  and  $Y \subseteq X$ , the two sets have the same elements and are said to be **equal**, which is expressed by

$$X = Y.$$

But when  $X \subseteq Y$  and  $X \neq Y$ , then  $X$  is called a **proper subset** of  $Y$ , and we use the more specific expression

$$X \subset Y.$$

The **intersection** of two sets  $X$  and  $Y$ , denoted by

$$X \cap Y,$$

consists of elements that are contained in both  $X$  and  $Y$ . The **union**

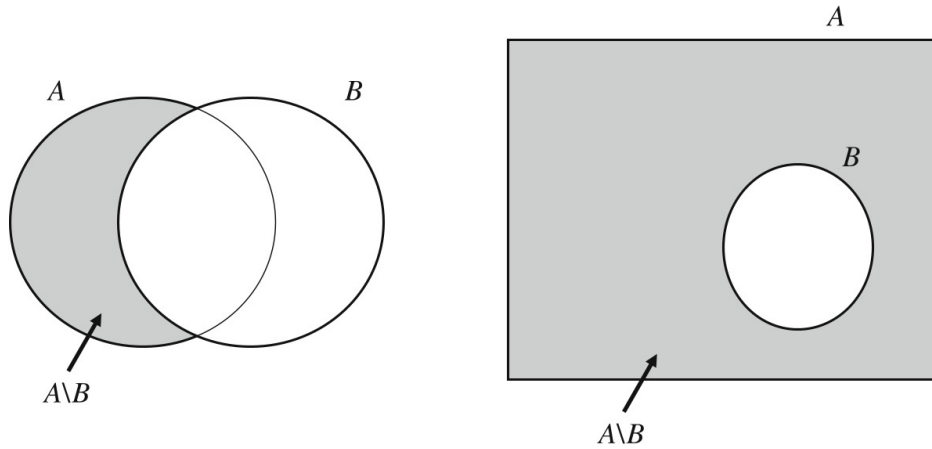
$$X \cup Y$$

consists of all the elements contained in either  $X$  or  $Y$ , including those contained in both  $X$  and  $Y$ . When the two sets  $X$  and  $Y$  have no element in common (i.e., when  $X \cap Y = \emptyset$ ),  $X$  and  $Y$  are said to be **disjoint**.

For two sets  $A$  and  $B$ , we define their **difference** by the set

$$\{x : x \in A, x \notin B\}$$

and denote it by  $A \setminus B$  (see Fig. 1.1). In particular, if  $A$  contains all the sets under discussion, we say that  $A$  is the **universal set** and  $A \setminus B$  is called the **complementary set** or **complement** of  $B$  in  $A$ .



**Fig. 1.1.** *Left:* The difference of two sets  $A$  and  $B$ . *Right:* The complementary set or complement of  $B$  in  $A$

### 1.1.2 Number Sets

Our abbreviations for fundamental number systems are given by

**N** : The set of all positive integers not including zero.

**Z** : The set of all integers.

**Q** : The set of all rational numbers.

**R** : The set of all real numbers.

**C** : The set of all complex numbers.

The symbol  $\mathbf{R}^n$  denotes an  $n$ -dimensional **Euclidean space** (see Sects. 4.1.3 and 19.2.3). Points in  $\mathbf{R}^n$  are denoted by bold face, say,  $\mathbf{x}$ ; the coordinates of  $\mathbf{x}$  are denoted by the ordered  $n$ -tuple  $(x_1, x_2, \dots, x_n)$ , where  $x_i \in \mathbf{R}$ . We also use the **extended real number** defined by

$$\overline{\mathbf{R}} = \mathbf{R} \cup \{-\infty, \infty\}.$$

### 1.1.3 Bounds

The precise terminology for bounds of real number sets follow. Meanwhile we assume  $S$  to be a set of real numbers.

♠ **Bounds of a set:**

1. A real number  $b$  such that  $x \leq b$  for all  $x \in S$  is called an **upper bound** of  $S$ .
2. A real number  $a$  such that  $x \geq a$  for all  $x \in S$  is called a **lower bound** of  $S$ .

Figure 1.2 illustrates the point. We say that a set  $S$  is **bounded above** or **bounded below** if it has an upper bound or a lower bound, respectively. In particular, when a set  $S$  is bounded above and below simultaneously, it is a **bounded set**. If a set  $S$  is not bounded, then it is said to be an **unbounded set**.

It follows from these definitions that if  $b$  is an upper bound of  $S$ , any number greater than  $b$  will also be an upper bound of  $S$ . Thus it makes sense to seek the *smallest* among such upper bounds. This is also the case for a lower bound of  $S$  if it is bounded below. In fact, the two extrema bounds, the smallest and the largest, are referred to by specific names as follows:

♠ **Least upper bound:**

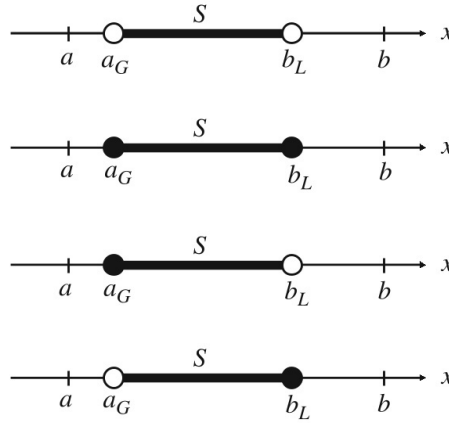
An element  $b \in \mathbf{R}$  is called the **least upper bound** (abbreviated by **l.u.b.**) or **supremum**, of  $S$  if

- (i)  $b$  is an upper bound of  $S$ , and
- (ii) there is no upper bound of  $S$  that is smaller than  $b$ .

♠ **Greatest lower bound:**

An element  $a \in \mathbf{R}$  is called the **greatest lower bound** (abbreviated by **g.l.b.**) or **infimum**, if

- (i)  $a$  is a lower bound of  $S$ , and
- (ii) there is no lower bound of  $S$  that is greater than  $a$ .



**Fig. 1.2.** In all the figures, the points  $a$  and  $b$  are lower and upper bounds of  $S$ , respectively. In particular, the point  $a_G$  is the greatest lower bound, and the  $b_L$  is the least upper bound

In symbols, the supremum and infimum of  $S$  are denote, respectively, by

$$\sup S \text{ and } \inf S.$$

We must emphasize the fact that the supremum and infimum of the set  $S$  may or may not belong to  $S$ . For instance, the set  $S = \{x : x < 1\}$  has the supremum 1, which it does not belong to  $S$ . Nevertheless, particularly when  $S$  is finite, we have

$$\sup S = \max S \text{ and } \inf S = \min S,$$

where  $\max S$  and  $\min S$  denote the **maximum** and **minimum** of  $S$ , respectively, both of which belong to  $S$ .

#### 1.1.4 Interval

When a set of real numbers is bounded above or below (or both), it is referred to as an **interval**; there are several classes of intervals as listed below.

- ♠ **Intervals:** Given a real variable  $x$ , the set of all values of  $x$  such that
1.  $a \leq x \leq b$  is a **closed interval**, denoted by  $[a, b]$ .
  2.  $a < x < b$  is a **bounded open interval**, denoted by  $(a, b)$ .
  3.  $a < x$  and  $x < b$  are **unbounded open intervals**, denoted by  $(a, \infty)$  and  $(-\infty, b)$ , respectively.

Sets of points  $\{x\}$  such that

$$a \leq x < b, a < x \leq b, a \leq x, x \leq b$$

may be referred to as **semiclosed intervals**; see Sect. 1.1.5 for more rigorous definitions. Every interval  $I_1$  contained in another interval  $I_2$  is a **subinterval** of  $I_2$ .

### 1.1.5 Neighborhood and Contact Point

The following is a preliminary definition that will be significant in the discussions on continuity and convergence properties of sets and functions.

♠ **Neighborhoods:**

Let  $x \in \mathbf{R}$ . A set  $V \subseteq \mathbf{R}$  is called a **neighborhood** of  $x$  if there is a number  $\varepsilon > 0$  such that

$$(x - \varepsilon, x + \varepsilon) \subseteq V.$$

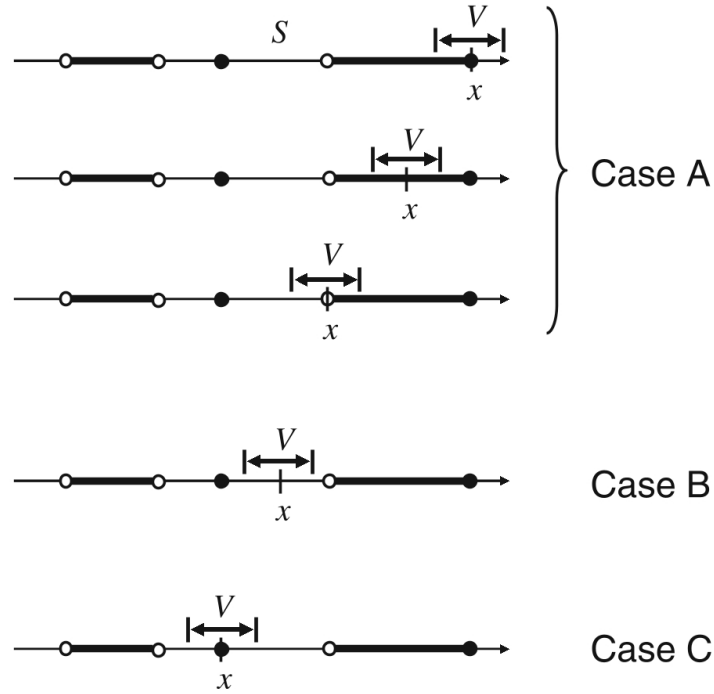
In line with the idea of neighborhoods, we introduce the following important concept (see Fig. 1.3):

♠ **Contact points:**

Assume a point  $x \in \mathbf{R}$  and a set  $S \subseteq \mathbf{R}$ . Then  $x$  is called a **contact point** of  $S$  if and only if *every* neighborhood of  $x$  contains at least one point of  $S$ .

**Remark.** A contact point of  $S$  may or may not belong to  $S$ . In contrast, a point  $x \in S$  is necessarily a contact point of  $S$ .

Obviously, every point of  $S$  is a contact point of  $S$ . In particular, when  $S$  is a single-element set given by  $S = \{x_0\}$  with  $x_0 \in \mathbf{R}$ , then  $x_0$  is a contact point of  $S$  since every neighborhood of  $x_0$  contains  $x_0$  itself. The collection of all contact points of a set  $S$  is called the **closure** of  $S$  and is denoted by  $[S]$ .



**Fig. 1.3.** Case A:  $x$  is a limit point (and thus a contact point) of  $S$ . Case B:  $x$  is *not* a contact point of  $S$ . Case C:  $x$  is an isolated point (and thus a contact point) of  $S$

Contact points can be classified as follows (see again Fig. 1.3):

♠ **Limit points:**

A contact point  $x \in \mathbf{R}$  is called a **limit point** of the set  $S \subseteq \mathbf{R}$  if and only if every neighborhood  $V$  of  $x$  contains a point of  $S$  *different from*  $x$ .

♠ **Isolated points:**

A contact point  $x$  is called an **isolated point** of  $S$  if and only if  $x$  has a neighborhood  $V$  in which  $x$  is the only point belonging to  $S$ .

In plain words, a limit point  $x$  is a point such that every interval  $(x - \varepsilon, x + \varepsilon)$  contains an infinite number of points, regardless of the smallness of  $\varepsilon$ . A limit point may be referred to as a **cluster point** or **accumulation point**, depending on the context. The symbol  $\hat{S}$  is commonly used to denote the set of limit points of  $S$ .

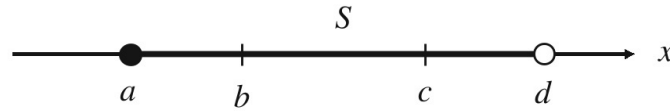
*Examples 1.* If  $S$  is the set of rational numbers in the interval  $[0, 1]$ , then every point of  $[0, 1]$ , rational or not, is a limit point of  $S$ .

**2.** The integer set  $\mathbf{Z}$  has no limit point; it has an infinite number of isolated points.

**3.** The origin is the limit point of the set  $\{1/m : m \in \mathbf{N}\}$ .

**Remark.** From the definition, a limit point of a set need not belong to the set. For instance,  $x = 1$  is the limit point of the set  $S = \{x : x \in \mathbf{R}, x > 1\}$ , but it does not belong to  $S$ . In contrast, an isolated point of  $S$  must lie in  $S$ .

Limit points are further divided into two classes. A limit point  $x$  of a set  $S$  is called an **interior point** of  $S$  if and only if  $x$  has a neighborhood  $V \subseteq S$ . Otherwise, it is called a **boundary point** of  $S$ . Figure 1.4 is a schematic illustration of the difference between interior and boundary points.



**Fig. 1.4.** All four points are limit points of  $S$ . Among them,  $b$  and  $c$  are interior points, whereas  $a$  and  $d$  are boundary points

### 1.1.6 Closed and Open Sets

Closed and open sets are defined in terms of the concepts of contact points and closure. Recall that a closure of  $S$ , denoted by  $[S]$ , is a set of all contact points of  $S$ , which is a union of the two sets: all limit and all isolated points of  $S$ .

♠ **Closed sets:**

A set  $S \subseteq \mathbf{R}$  is **closed** if  $[S] = S$ , i.e., if  $S$  coincides with its own closure.

♠ **Open sets:**

A set  $S \subseteq \mathbf{R}$  is **open** if  $S$  consists entirely of its interior points and has no boundary points.

It follows intuitively that a set  $S \subseteq \mathbf{R}$  is open if and only if its **complementary set** is closed. The proof is given in Exercise 4 in this chapter. Note that the condition  $[S] \neq S$  is inconclusive as to whether  $S$  is open or not.

*Examples* 1. Every single-element set  $S = \{x_0\}$  with  $x_0 \in \mathbf{R}$  is closed since  $[S] = S$ .

2. Every set consisting of a finite number of points is closed.
3. For any real number  $x$ , the set  $\mathbf{R} \setminus \{x\}$  is open since  $\{x\}$  is closed.
4. The intervals  $[a, b]$ ,  $[a, \infty)$ , and  $(-\infty, b]$  are all closed, which is proven by considering their closures.
5. The interval  $[a, b)$  is neither closed nor open. In fact, it is not closed since it excludes its boundary point  $b$  and it is not open since it contains its boundary point  $a$ .

## Exercises

1. Give the supremum and infimum of each of the following sets:

- (1)  $S = \{x : 0 \leq x \leq 5\}$ .
- (2)  $S = \{x : x \in \mathbf{Q} \text{ and } x^2 < 2\}$ .
- (3)  $S = \{x : x = 3 + \frac{1}{n^2}, n \in \mathbf{N}\}$ .

**Solution:** (1)  $\sup S = 5$ ,  $\inf S = 0$ . (2)  $\sup S = \sqrt{2}$ ,  $\inf S = 0$ .  
 (3)  $\sup S = 4$ ,  $\inf S = 3$ . ♣

2. Suppose  $S$  to be any of the intervals:  $(a, b)$ ,  $[a, b)$ ,  $(a, b]$ , or  $[a, b]$ . Show that

$$\sup S = b, \quad \inf S = a.$$

**Solution:** Take  $S = (a, b)$ . Since  $x \leq b$  for all  $x \in S$ ,  $b$  serves as one of upper bounds of  $S$ . We show that  $b$  is surely the *least* upper bound. To see this, we first assume that  $u$  is another upper bound of  $S$  such that  $u < b$ ; then  $a < u < (u + b)/2 < b$ . This implies that

$$\frac{u + b}{2} \in S \quad \text{and} \quad u < \frac{u + b}{2},$$

which contradicts the assumption that  $u$  is an upper bound of  $S$ . Hence,  $u \geq b$ ; i.e., any upper bound other than  $b$  must be larger than  $b$ . We thus conclude that  $b = \sup S$ . The proof is similar for the other three cases. ♣

3. Show that the set of integers has no limit point, i.e.,  $\hat{\mathbf{Z}} = \emptyset$ .

**Solution:** Take any  $x \in \mathbf{Z}$ , and let  $\varepsilon = \min\{|n - x| : n \in \mathbf{Z}\}$ . The interval  $(x - \varepsilon, x + \varepsilon)$  contains no integers other than  $x$ ; hence,  $x \notin \hat{\mathbf{Z}}$ . Since this is the case for any  $x \in \mathbf{Z}$ , we conclude that  $\mathbf{Z}$  is totally composed of isolated points. ♣

4. Show that a set  $S \subseteq \mathbf{R}$  is open if and only if its complementary set  $\mathbf{R} \setminus S$  is closed.

**Solution:** If  $S$  is open, then every point  $x \in S$  has a neighborhood contained in  $S$ . Therefore no point  $x \in S$  can be a contact point of  $\mathbf{R} \setminus S$ . In other words, if  $x$  is a contact point of  $\mathbf{R} \setminus S$ , then  $x \in \mathbf{R} \setminus S$ , i.e.,  $\mathbf{R} \setminus S$  is closed.

Conversely, if  $\mathbf{R} \setminus S$  is closed, then any point  $x \in S$  must have a neighborhood contained in  $S$ , since otherwise every neighborhood

of  $x$  would contain points of  $\mathbf{R} \setminus S$ , i.e.,  $x$  would be a contact point of  $\mathbf{R} \setminus S$  not in  $\mathbf{R} \setminus S$ . Therefore  $S$  is open. ♣

## 1.2 Conditional Statements

Phrases such as *if... then...*, and *... if and only if ...* are frequently used to connect **simple statements** that can be described as either **true** or **false**. For the sake of typographical convenience, there are conventional logical symbols for representing such phrases.

Suppose  $P$  and  $Q$  are two different statements. The compound statements

if  $P$  then  $Q$

and

$P$  implies  $Q$

mean that if  $P$  is true then  $Q$  is true. This is written symbolically as

$$P \Rightarrow Q. \quad (1.1)$$

We say that

$P$  is a **sufficient condition** for  $Q$

or

$Q$  is a **necessary condition** for  $P$ .

In the above context,  $P$  stands for the hypothesis or assumption, and  $Q$  is the conclusion.

**Remark.** To prove the implication (1.1) in actual problems, it suffices to exclude the possibility that  $P$  is true and  $Q$  is false. This may be done in one of three ways.

1. Assume that  $P$  is true and prove that  $Q$  is true (**direct proof**).
2. Assume that  $Q$  is false and prove that  $P$  is false (**contrapositive proof**).
3. Assume that  $P$  is true and  $Q$  is false, and then prove that this leads to a contradiction (**proof by contradiction**).

When  $P$  implies  $Q$  and  $Q$  implies  $P$ , we abbreviate this to

$$P \Longleftrightarrow Q,$$

and we say that

$P$  is **equivalent** to  $Q$

or, more commonly,

$P$  **if and only if**  $Q$ .

This also means that  $P$  is a **necessary and sufficient condition** for  $Q$ .

*Examples* Observe that

$$x = 1 \Rightarrow x^2 = 1 \text{ and } x = -1 \Rightarrow x^2 = 1.$$

Conversely, we see that

$$x^2 = 1 \Rightarrow x = -1 \text{ or } 1.$$

Therefore, we conclude that

$$x^2 = 1 \iff x \in \{-1, 1\}.$$

## 1.3 Order of Magnitude

### 1.3.1 Symbols $O$ , $o$ , and $\sim$

We use the notations  $O$ ,  $o$ , and  $\sim$  to express orders of magnitude. To explain their use, we consider the behavior of functions  $f(x)$  and  $g(x)$  in a neighborhood of a point  $x_0$ .

**1.** We write

$$f(x) = O(g(x)), \quad x \rightarrow x_0$$

if there exists a positive constant  $A$  such that

$$|f(x)| \leq A|g(x)|$$

for all values of  $x$  in some neighborhood of  $x_0$ .

**2.** We write

$$f(x) = o(g(x)), \quad x \rightarrow x_0$$

if

$$\lim_{x \rightarrow x_0} \left| \frac{f(x)}{g(x)} \right| = 0.$$

**3.** We write

$$f(x) \sim g(x), \quad x \rightarrow x_0$$

if

$$\lim_{x \rightarrow x_0} \frac{f(x)}{g(x)} = 1.$$

In addition to the formal definitions above, we summarize the actual meaning of these symbols:

1.  $f(x) = O(g(x))$  means that  $f(x)$  does *not* grow *faster* than  $g(x)$  as  $x \rightarrow x_0$ .
2.  $f(x) = o(g(x))$  means that  $f(x)$  grows more *slowly* than  $g(x)$  as  $x \rightarrow x_0$ .
3.  $f(x) \sim g(x)$  means that  $f(x)$  and  $g(x)$  grow *at the same rate* as  $x \rightarrow x_0$ .

We occasionally employ the symbols

$$f(x) = O(1) \text{ as } x \rightarrow x_0.$$

This simply means that  $f(x)$  is bounded on the order of 1. The symbol

$$f(x) = o(1) \text{ as } x \rightarrow x_0$$

means that  $f(x)$  approaches zero as  $x \rightarrow x_0$ .

*Examples* The relations **1–3** below hold for  $x \rightarrow \infty$ .

1.  $\frac{1}{1+x^2} = O\left(\frac{1}{x^2}\right), \quad \frac{1}{1+x^2} = o\left(\frac{1}{x}\right), \quad \frac{1}{1+x^2} \sim \frac{1}{x^2}.$
2.  $\frac{1}{1+x^2} = \frac{1}{x^2} + O\left(\frac{1}{x^4}\right), \quad \frac{1}{1+x^2} = \frac{1}{x^2} + o\left(\frac{1}{x^2}\right).$
3.  $\sqrt{x^2+1} = x + O\left(\frac{1}{x}\right), \quad \sqrt{x^2+1} = x + o(1), \quad \sqrt{x^2+1} \sim x.$

The following hold for  $x \rightarrow 0$ :

4.  $\sin x = O(1), \quad \sin x \sim x, \quad \cos x = 1 + O(x^2).$

### 1.3.2 Asymptotic Behavior

Asymptotic behavior of  $f(x)$  as  $x \rightarrow a$  can be quantified by using the powers of  $(x-a)$  as comparison functions. As an example, suppose that a function  $f(x)$  satisfies the relation

$$f(x) = O((x-a)^p) \text{ for } x \rightarrow a \tag{1.2}$$

for some real number  $p = p_0$ . Then, the relation (1.2) clearly holds for all  $p$  for  $p \leq p_0$ , and it may or may not hold for some  $p$  if  $p \geq p_0$ . Thus we can define the **supremum** of such  $p$ 's that satisfy (1.2), and denote it by  $q$ , i.e.,

$$q = \sup\{p \mid f(x) = O((x-a)^p)\}. \tag{1.3}$$

In this case, we say that  $f$  **vanishes** at  $x = a$  to **order**  $q$ . The quantity  $q$  defined by (1.3) is useful for describing the asymptotic behavior of  $f(x)$  in the vicinity of  $x = a$ .

*Remark.* Note that (1.3) itself does not imply that

$$f(x) = O((x-a)^q), \quad x \rightarrow a.$$

For instance, the function  $f(x) = \log x$  defined within the interval  $(0, 1)$  yields  $q = 0$ , since for  $x \rightarrow 0$ ,

$$\log x \begin{cases} = O(x^p) & p < 0, \\ \neq O(x^p) & p > 0. \end{cases}$$

But it is obvious that  $\log x \neq O(1)$ .

## 1.4 Values of Indeterminate Forms

### 1.4.1 l'Hôpital's Rule

A function  $f(x)$  of the form  $u(x)/v(x)$  is not defined for  $x = a$  if  $f(a)$  takes the form  $0/0$ . Still, if the limit  $\lim_{x \rightarrow a} f(x)$  exists, then it is often desirable to define  $f(a) \equiv \lim_{x \rightarrow a} f(x)$ . In such a case, the value of the limit can be evaluated by using the following theorem:

#### ♠ l'Hôpital's rule:

Let  $u(a) = v(a) = 0$ . If there exists a neighborhood of  $x = a$  such that

- (i)  $v(x) \neq 0$  except for  $x = a$ , and
- (ii)  $u'(x)$  and  $v'(x)$  exist and do not vanish simultaneously, then,

$$\lim_{x \rightarrow a} \frac{u(x)}{v(x)} = \lim_{x \rightarrow a} \frac{u'(x)}{v'(x)}$$

whenever the limit on the right exists.

For the proof of the theorem, see Exercise **3** in Sect. 8.1.

*Remark.* If  $u'(x)/v'(x)$  is itself an indeterminate form, the above method may be applied to  $u'(x)/v'(x)$  in turn, so that

$$\lim_{x \rightarrow a} \frac{u(x)}{v(x)} = \lim_{x \rightarrow a} \frac{u'(x)}{v'(x)} = \lim_{x \rightarrow a} \frac{u''(x)}{v''(x)}.$$

If necessary, this process may be continued.

### 1.4.2 Several Examples

In the following, we show several examples of indeterminate forms other than the form of  $0/0$  previously discussed. Often functions  $f(x)$  of the forms  $u(x)v(x)$ ,  $[u(x)]^{v(x)}$ , and  $u(x) - v(x)$  can be reduced to the form  $p(x)/q(x)$  with the aid of the following relations:

$$\begin{aligned} u(x)v(x) &= \frac{u(x)}{1/v(x)} = \frac{v(x)}{1/u(x)}, \\ [u(x)]^{v(x)} &= e^{g(x)}, \text{ where } g(x) = \frac{\log u(x)}{1/v(x)} = \frac{\log v(x)}{1/u(x)}, \\ u(x) - v(x) &= \frac{\frac{1}{v(x)} - \frac{1}{u(x)}}{\frac{1}{u(x)} \frac{1}{v(x)}} = \log h(x), \text{ where } h(x) = \frac{e^{u(x)}}{e^{v(x)}}. \end{aligned}$$

After the reduction, the l'Hôpital method given in Sect. 1.4.1 becomes applicable.



## Part I

---

## Real Analysis



## Real Sequences and Series

---

**Abstract** In this chapter, we deal with the fundamental properties of sequences and series of real numbers. We place particular emphasis on the concept of “convergence,” a thorough understanding of which is important for the study of the various branches of mathematical physics that we are concerned with subsequent chapters.

### 2.1 Sequences of Real Numbers

#### 2.1.1 Convergence of a Sequence

This section describes the fundamental definitions and ideas associated with **sequences of real numbers** (called **real sequences**). We must emphasize that the sequence

$$(x_n : n \in \mathbf{N})$$

is not the same as the set

$$\{x_n : n \in \mathbf{N}\}.$$

In fact, the former is the *ordered* list of  $x_n$ , some of which may be repeated, whereas the latter is merely the defining *range* of  $x_n$ . For instance, the constant sequence  $x_n = 1$  is denoted by  $(1, 1, 1, \dots)$ , whereas the set  $\{1\}$  contains only one element.

We start with a precise definition of the convergence of a real sequence, which is an initial and crucial step for various branches of mathematics.

♠ **Convergence of a real sequence:**

A real sequence  $(x_n)$  is said to be **convergent** if there exists a real number  $x$  with the following property: For every  $\varepsilon > 0$ , there is an integer  $N$  such that

$$n \geq N \Rightarrow |x_n - x| < \varepsilon. \quad (2.1)$$

We must emphasize that the magnitude of  $\varepsilon$  is arbitrary. No matter how small an  $\varepsilon$  we choose, it must always be possible to find a number  $N$  that will increase as  $\varepsilon$  decreases.

**Remark.** In the language of **neighborhoods**, the above definition is stated as follows: *The sequence  $(x_n)$  converges to  $x$  if every neighborhood of  $x$  contains all but a finite number of elements of the sequence.*

When  $(x_n)$  is convergent, the number  $x$  specified in this definition is called a **limit** of the sequence  $(x_n)$ , and we say that  $x_n$  converges to  $x$ . This is expressed symbolically by writing

$$\lim_{n \rightarrow \infty} x_n = x,$$

or simply by

$$x_n \rightarrow x.$$

If  $(x_n)$  is not convergent, it is called **divergent**.

**Remark.** The limit  $x$  may or may not belong to  $(x_n)$ ; this situation is similar to the case of the **limit point** of a set of real numbers discussed in Sect. 1.1.5.

An example in which  $x = \lim x_n$  but  $x \neq x_n$  for any  $n$  is given below.

*Examples* Suppose that a sequence  $(x_n)$  consisting of rational numbers is defined by

$$(x_n) = (3.1, 3.14, 3.142, \dots, x_n, \dots),$$

where  $x_n \in \mathbf{Q}$  is a rational number to  $n$  decimal places close to  $\pi$ . Since the difference  $|x_n - \pi|$  is less than  $10^{-n}$ , it is possible to find an  $N$  for any  $\varepsilon > 0$  such that

$$n \geq N \Rightarrow |x_n - \pi| < \varepsilon.$$

This means that

$$\lim_{n \rightarrow \infty} x_n = \pi.$$

However, as the limit,  $\pi$ , is an irrational number it is not in  $\mathbf{Q}$ .

**Remark.** The above example indicates that only a restricted class of convergent sequences has a limit in the same sequence.

### 2.1.2 Bounded Sequences

In the remainder of this section, we present several fundamental concepts associated with real sequences. We start with the boundedness properties of sequences.

♠ **Bounded sequences:**

A real sequence  $(x_n)$  is said to be **bounded** if there is a positive number  $M$  such that

$$|x_n| \leq M \text{ for all } n \in \mathbf{N}.$$

The following is an important relation between convergence and boundedness of a real sequence:

♠ **Theorem:**

If a sequence is convergent, then it is bounded.

*Proof* Suppose that  $x_n \rightarrow x$ . If we choose  $\varepsilon = 1$  in (2.1), there exists an integer  $N$  such that

$$|x_n - x| < 1 \text{ for all } n \geq N.$$

Since  $|x_n| - |x| \leq |x_n - x|$ , it follows that

$$|x_n| < 1 + |x| \text{ for all } n \geq N.$$

Setting  $M = \max\{|x_1|, |x_2|, \dots, |x_{N-1}|, 1 + |x|\}$  yields

$$|x_n| < M \text{ for all } n \in \mathbf{N},$$

which means that  $(x_n)$  is bounded. ♣

*Remark.* Observe that the converse of the theorem is false. In fact, the sequence

$$(1, -1, 1, -1, \dots, (-1)^n, \dots)$$

is divergent, although it is bounded.

### 2.1.3 Monotonic Sequences

Another important concept in connection with real sequences is monotonicity, defined as follows:

♠ **Monotonic sequences:**

A sequence  $(x_n)$  is said to be

1. **increasing** (or monotonically increasing) if  $x_{n+1} \geq x_n$  for all  $n \in \mathbf{N}$ ,
2. **strictly increasing** if  $x_{n+1} > x_n$  for all  $n \in \mathbf{N}$ ,

- 3. decreasing** (or monotonically decreasing) if  $x_{n+1} \leq x_n$  for all  $n \in \mathbf{N}$ , and
- 4. strictly decreasing** if  $x_{n+1} < x_n$  for all  $n \in \mathbf{N}$ .

These four kinds of sequences are collectively known as **monotonic sequences**. Note that a sequence  $(x_n)$  is increasing if and only if  $(-x_n)$  is decreasing. Thus, the properties of monotonic sequences can be fully investigated by restricting ourselves solely to increasing (or decreasing) sequences.

Once a sequence assumes monotonic properties, its convergence is determined only by its boundedness, as stated below.

♠ **Theorem:**

A monotonic sequence is convergent if and only if it is bounded. More specifically,

- (i) If  $(x_n)$  is increasing and bounded above, then its limit is given by

$$\lim_{n \rightarrow \infty} x_n = \sup x_n.$$

- (ii) If  $(x_n)$  is decreasing and bounded below, then

$$\lim_{n \rightarrow \infty} x_n = \inf x_n.$$

*Proof* If  $(x_n)$  is convergent, then it must be bounded as proven earlier (see Sect. 2.1.2). Now we consider the converses for cases (i) and (ii).

- (i) Assume  $(x_n)$  is increasing and bounded. The set  $S = \{x_n\}$  will then have the supremum denoted by  $\sup S = x$ . By the definition of the supremum, for arbitrary small  $\varepsilon > 0$  there is an  $x_N \in S$  such that

$$x_N > x - \varepsilon. \quad (2.2)$$

Since  $x_n$  is increasing, we obtain

$$x_n \geq x_N \text{ for all } n \geq N. \quad (2.3)$$

Moreover, since  $x$  is the supremum of  $S$ , we have

$$x \geq x_n \text{ for all } n \in \mathbf{N}. \quad (2.4)$$

From (2.2), (2.3), and (2.4), we arrive at

$$|x_n - x| = x - x_n \leq x - x_N < \varepsilon \text{ for all } n \geq N,$$

which gives us the desired conclusion, i.e.,

$$\lim_{n \rightarrow \infty} x_n = x = \sup S.$$

- (ii) If  $(x_n)$  is decreasing and bounded, then  $(-x_n)$  is increasing and bounded. Hence, from (i), we have

$$\lim_{n \rightarrow \infty} (-x_n) = \sup(-S).$$

Since  $\sup(-S) = -\inf S$ , it follows that

$$\lim_{n \rightarrow \infty} x_n = \inf S. \quad \clubsuit$$

#### 2.1.4 Limit Superior and Limit Inferior

We close this section by introducing two specific limits on any bounded sequence. Let  $(x_n)$  be a bounded sequence and define two sequences  $(y_n)$  and  $(z_n)$  as follows:

$$\begin{aligned} y_n &= \sup\{x_k : k \geq n\}, \\ z_n &= \inf\{x_k : k \geq n\}. \end{aligned} \tag{2.5}$$

Note that  $y_n$  and  $z_n$  differ, respectively, from  $\sup\{x_n\}$  and  $\inf\{x_n\}$ . It follows from (2.5) that

$$y_1 = \sup\{x_k : k \geq 1\} \geq y_2 = \sup\{x_k : k \geq 2\} \geq y_3 \cdots,$$

which means that the sequence  $(y_n)$  is monotonically decreasing and bounded below by  $\inf x_n$ . Thus in view of the theorem in Sect. 2.1.3, the sequence  $(y_n)$  must be convergent. The limit of  $(y_n)$  is called the **limit superior** or the **upper limit** of  $(x_n)$  and is denoted by

$$\limsup_{n \rightarrow \infty} x_n \quad (\text{or } \overline{\lim} x_n).$$

Likewise, since  $(z_n)$  is increasing and bounded above by  $\sup x_n$ , it possesses the limit known as the **limit inferior** or **lower limit** of  $x_n$  denoted by

$$\liminf_{n \rightarrow \infty} x_n \quad (\text{or } \underline{\lim} x_n).$$

In terms of the two specific limits, we can say that a bounded sequence  $(x_n)$  converges if and only if

$$\lim_{n \rightarrow \infty} x_n = \limsup_{n \rightarrow \infty} x_n = \liminf_{n \rightarrow \infty} x_n.$$

(A proof will be given in Exercise 4 in Sect. 2.1.4.) Note that by definition, it readily follows that

$$\begin{aligned} \limsup_{n \rightarrow \infty} x_n &\geq \liminf_{n \rightarrow \infty} x_n, \\ \limsup_{n \rightarrow \infty} (-x_n) &= -\liminf_{n \rightarrow \infty} x_n. \end{aligned}$$

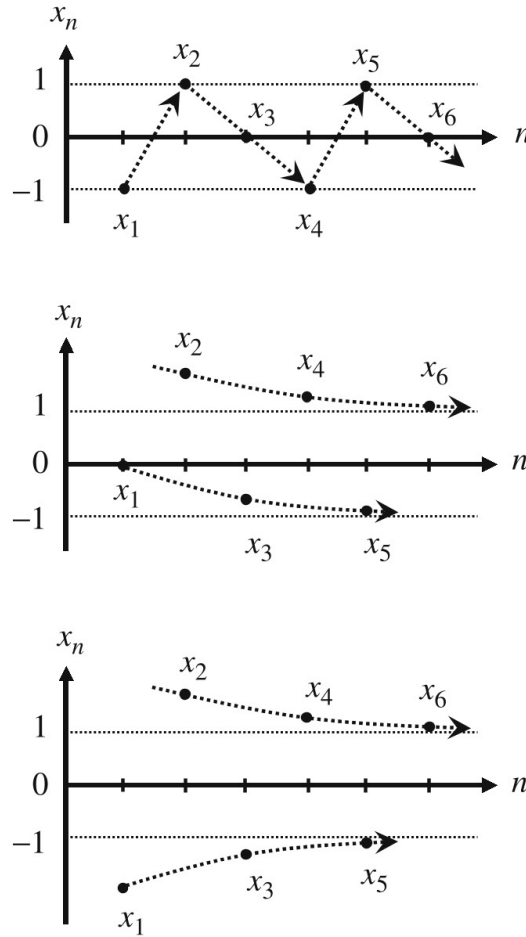
*Examples* 1.  $x_n = (-1)^n \Rightarrow \limsup_{n \rightarrow \infty} x_n = 1, \liminf_{n \rightarrow \infty} x_n = -1.$

2.  $x_n = (-1)^n + \frac{1}{n} \Rightarrow \limsup_{n \rightarrow \infty} x_n = 1, \liminf_{n \rightarrow \infty} x_n = -1.$

3.  $x_{2n} = 1 + \frac{(-1)^n}{n}, x_{2n-1} = \frac{(-1)^n}{n}, \Rightarrow \limsup_{n \rightarrow \infty} x_n = 1, \liminf_{n \rightarrow \infty} x_n = 0.$

4.  $(x_n) = (2, 0, -2, 2, 0, -2, \dots) \Rightarrow \limsup_{n \rightarrow \infty} x_n = 2, \liminf_{n \rightarrow \infty} x_n = -2.$

The four cases noted above are illustrated schematically in Fig. 2.1. All the sequences  $(x_n)$  are not convergent and thus the limit  $\lim_{n \rightarrow \infty} x_n$  does not exist. This fact clarifies the crucial difference between  $\lim_{n \rightarrow \infty} x_n$  and  $\limsup_{n \rightarrow \infty} x_n$  (or  $\liminf_{n \rightarrow \infty} x_n$ ).



**Fig. 2.1.** All the sequences of  $\{x_n\}$  in the figures do not converge, but they all possess  $\limsup_{n \rightarrow \infty} x_n = 1$  and  $\liminf_{n \rightarrow \infty} x_n = -1$

The limit superior of  $x_n$  has the following features and similar features are found for the limit inferior.

**♠ Theorem:**

1. For any small  $\varepsilon > 0$ , we can find an  $N$  such that

$$n > N \Rightarrow x_n < \limsup_{n \rightarrow \infty} x_n + \varepsilon.$$

2. For any small  $\varepsilon > 0$ , there are an infinite number of terms of  $x_n$  such that

$$\limsup_{n \rightarrow \infty} x_n - \varepsilon < x_n.$$

*Proof* 1. Recall that  $\limsup_{n \rightarrow \infty} x_n = \lim_{n \rightarrow \infty} y_n$ , where  $y_n$  is defined in (2.5). For any  $\varepsilon > 0$ , there is an integer  $N$  such that

$$n > N \Rightarrow \limsup_{n \rightarrow \infty} x_n - \varepsilon < y_n < \limsup_{n \rightarrow \infty} x_n + \varepsilon.$$

Since  $y_n \geq x_n$  for all  $n$ , we have

$$n > N \Rightarrow x_n < \limsup_{n \rightarrow \infty} x_n + \varepsilon. \quad \clubsuit$$

2. Suppose that there is an integer  $m$  such that

$$n > m \Rightarrow \limsup_{n \rightarrow \infty} x_n - \varepsilon \geq x_n.$$

Then for all  $k \geq n > m$ , we have

$$x_k \leq \limsup_{n \rightarrow \infty} x_n - \varepsilon,$$

which means that

$$y_n \leq \limsup_{n \rightarrow \infty} x_n - \varepsilon \text{ for all } n > m.$$

In the limit of  $n \rightarrow \infty$ , we find a contradiction such that

$$\limsup_{n \rightarrow \infty} x_n \leq \limsup_{n \rightarrow \infty} x_n - \varepsilon.$$

This completes the proof.  $\clubsuit$

**Exercises**

1. Prove that if the sequence  $(x_n)$  is convergent, then its limit is unique.

**Solution:** Let  $x = \lim x_n$  and  $y = \lim x_n$  with the assumption  $x \neq y$ . Then we can find a neighborhood  $V_1$  of  $x$  and a neighborhood  $V_2$  of  $y$  such that  $V_1 \cap V_2 = \emptyset$ . For example, take  $V_1 = (x - \varepsilon, x + \varepsilon)$  and  $V_2 = (y - \varepsilon, y + \varepsilon)$ , where  $\varepsilon = |x - y|/2$ . Since  $x_n \rightarrow x$ , all but a finite number of terms of the sequence lie in  $V_1$ . Similarly, since  $y_n \rightarrow y$ , all but a finite number of its terms also lie in  $V_2$ . However, these results contradict the fact that  $V_1 \cap V_2 = \emptyset$ , which means that the limit of a sequence should be unique. ♣

2. If  $x_n \rightarrow x \neq 0$ , then there is a positive number  $A$  and an integer  $N$  such that  $n > N \Rightarrow |x_n| > A$ . Prove it.

**Solution:** Let  $\varepsilon = |x|/2$ , which is a positive number. Hence, there is an integer  $N$  such that  $n > N \Rightarrow |x_n - x| < \varepsilon \Rightarrow ||x_n| - |x|| < \varepsilon$ . Consequently,  $|x| - \varepsilon < |x_n| < |x| + \varepsilon$  for all  $n \geq N$ . From the left-hand inequality, we see that  $|x_n| > |x|/2$ , and we can take  $M = |x|/2$  to complete the proof. ♣

3. Prove that the sequence  $x_n = [1 + (1/n)]^n$  is convergent.

**Solution:** The proof is completed by observing that the sequence is monotonically increasing and bounded. To see this, we use the **binomial theorem**, which gives

$$\begin{aligned} x_n &= \sum_{k=0}^n {}^nC_{n-k} \frac{1}{n^k} \\ &= 1 + 1 + \frac{1}{2!} \left(1 - \frac{1}{n}\right) + \frac{1}{3!} \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) + \cdots \\ &\quad + \frac{1}{n!} \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{n-1}{n}\right). \end{aligned}$$

Likewise we have

$$\begin{aligned} x_{n+1} &= 1 + 1 + \frac{1}{2!} \left(1 - \frac{1}{n+1}\right) + \frac{1}{3!} \left(1 - \frac{1}{n+1}\right) \left(1 - \frac{2}{n+1}\right) + \cdots \\ &\quad + \frac{1}{(n+1)!} \left(1 - \frac{1}{n+1}\right) \left(1 - \frac{2}{n+1}\right) \cdots \left(1 - \frac{n}{n+1}\right). \end{aligned}$$

Comparing these expressions for  $x_n$  and  $x_{n+1}$ , we see that every term in  $x_n$  is no more than corresponding term in  $x_{n+1}$ . In addition,  $x_{n+1}$  has an extra positive term. We thus conclude that  $x_{n+1} \geq x_n$  for all  $n \in \mathbf{N}$ , which means that the sequence  $(x_n)$  is monotonically increasing.

We next prove boundedness. For every  $n \in \mathbf{N}$ , we have  $x_n < \sum_{k=0}^n (1/k!)$ . Using the inequality  $2^{n-1} \leq n!$  for  $n \geq 1$  (which can be easily seen by taking the logarithm of both sides), we obtain

$$x_n < 1 + \sum_{k=1}^n \frac{1}{2^{k-1}} = 1 + \frac{1 - (1/2)^n}{1 - (1/2)} < 3.$$

Thus  $(x_n)$  is bounded above by 3. Thus, view of the theorem in Sect. 2.1.3, the sequence is convergent. ♣

4. Denote  $\bar{x} = \limsup x_n$  and  $\underline{x} = \liminf x_n$ . Prove that a sequence  $(x_n)$  converges to  $x$  if and only if  $x = \bar{x} = \underline{x}$ .

**Solution:** In view of the theorem in Sect. 2.1.4, it follows that  $(-\infty, \bar{x} + \varepsilon)$  contains all but a finite number of terms of  $(x_n)$ . The same property applied to  $(-x_n)$  implies that  $(\underline{x} - \varepsilon, \infty)$  contains all but a finite number of such terms. If  $x = \bar{x} = \underline{x}$ , then  $(x - \varepsilon, x + \varepsilon)$  contains all but a finite number of terms of  $(x_n)$ . This is the assertion that  $x_n \rightarrow x$ .

Now suppose that  $x_n \rightarrow x$ . For any  $\varepsilon > 0$ , there is an integer  $N$  such that  $n > N \Rightarrow x_n < x + \varepsilon \Rightarrow y_n \leq x + \varepsilon$ , where  $y_n = \sup\{x_k : k \geq n\}$ , as was introduced in (2.5). Hence,  $\bar{x} \leq x + \varepsilon$ . Since  $\varepsilon > 0$  is arbitrary, we obtain  $\bar{x} \leq x$ . Working with the sequence  $(-x_n)$ , whose limit is  $-x$ , following same procedure, we get  $\underline{x} \geq x$ . Since  $\underline{x} \leq \bar{x}$ , we conclude that  $x = \bar{x} = \underline{x}$ . ♣

## 2.2 Cauchy Criterion for Real Sequences

### 2.2.1 Cauchy Sequence

To test the convergence of a general (nonmonotonic) real sequence, we have thus far only the original definition given in Sect. 2.1.1 to rely on; in that case we must first have a candidate for the limit of the sequence in question before we can examine its convergence. Needless to say, it is more convenient if we can determine the convergence property of a sequence without having to guess its limit. This is achieved by applying the so-called **Cauchy criterion**, which plays a central role in developing the fundamentals of real analysis.

To begin with, we present a preliminary notion for subsequent discussions.

#### ♠ Cauchy sequence:

The sequence  $(x_n)$  is called a **Cauchy sequence** (or **fundamental sequence**) if for every positive number  $\varepsilon$ , there is a positive integer  $N$  such that

$$m, n > N \Rightarrow |x_n - x_m| < \varepsilon. \quad (2.6)$$

This means that in every Cauchy sequence, the terms can be as close to one another as we like. This feature of Cauchy sequences is expected to hold for any convergent sequence, since the terms of a convergent sequence have to approach each other as they approach a common limit. This conjecture is ensured in part by the following theorem.

**♠ Theorem:**

If a sequence  $(x_n)$  is convergent, then it is a Cauchy sequence.

*Proof* Suppose  $\lim x_n = x$  and  $\varepsilon$  is any positive number. From hypothesis, there exists a positive integer  $N$  such that

$$n > N \Rightarrow |x_n - x| < \frac{\varepsilon}{2}.$$

Now if we take  $m, n \geq N$ , then

$$|x_n - x| < \frac{\varepsilon}{2} \text{ and } |x_m - x| < \frac{\varepsilon}{2}.$$

It thus follows that

$$|x_n - x_m| \leq |x_m - x| + |x_n - x| < \varepsilon,$$

which means that  $(x_n)$  is a Cauchy sequence. ♣

This theorem naturally gives rise to a question as to whether converse true. In other words, we would like to know whether all Cauchy sequences are convergent or not. The answer is exactly what the Cauchy criterion states, as we prove in the next subsection.

**2.2.2 Cauchy Criterion**

The following is one of the fundamental theorems of real sequences.

**♠ Cauchy criterion:**

A sequence of real numbers is convergent if and only if it is a Cauchy sequence.

Bear in mind that the validity of this criterion was partly proven by demonstrating the previous theorem (see Sect. 2.2.1). Hence, in order to complete the proof of the criterion, we need only prove that *every Cauchy sequence is convergent*. The following serves as a lemma for developing the proof.

**♠ Bolzano – Weierstrass theorem:**

Every infinite and bounded sequence of real numbers has at least one limit point in  $\mathbf{R}$ . (The proof is given in Appendix A.)

We are now ready to prove that *every Cauchy sequence is convergent*.

**Proof (of the Cauchy criterion):** Let  $(x_n)$  be a Cauchy sequence and  $S = \{x_n : n \in \mathbf{N}\}$ . We consider two cases in turn: **(i)** the set  $S$  is finite, and **(ii)**  $S$  is infinite.

- (i)** It follows from the hypothesis that given  $\varepsilon > 0$ , there is an integer  $N$  such that

$$m, n > N \Rightarrow |x_n - x_m| < \varepsilon. \quad (2.7)$$

Since  $S$  is finite, one of the terms of the sequence  $(x_n)$ , say  $x$ , should be repeated infinitely often in order to satisfy (2.7). This implies the existence of an  $m > N$  such that  $x_m = x$ . Hence, we have

$$n > N \Rightarrow |x_n - x| < \varepsilon,$$

which means that  $x_n \rightarrow x$ .

- (ii)** Next we consider the case that  $S$  is infinite. It can be shown that every Cauchy sequence is bounded (see Exercise 1). Hence, in view of the Bolzano – Weierstrass theorem, the sequence  $(x_n)$  necessarily has a limit point  $x$ . We shall prove that  $x_n \rightarrow x$ . Given  $\varepsilon > 0$ , there is an integer  $N$  such that

$$m, n > N \Rightarrow |x_n - x_m| < \varepsilon.$$

From the definition of a limit point, we see that the interval  $(x - \varepsilon, x + \varepsilon)$  contains an infinite number of terms of the sequence  $(x_n)$ . Hence, there is an  $m \geq N$  such that  $x_m \in (x - \varepsilon, x + \varepsilon)$ , i.e., such that  $|x_n - x_m| < \varepsilon$ . Now, if  $n \geq N$ , then

$$|x_n - x| \leq |x_n - x_m| + |x_m - x| < \varepsilon + \varepsilon = 2\varepsilon,$$

which proves  $x_n \rightarrow x$ .

The results for **(i)** and **(ii)** shown above indicate that *every Cauchy sequence (finite and infinite) is convergent*. Recall again that its converse, *every convergent sequence is a Cauchy sequence*, was proven earlier in Sect. 2.2.1. This completes the proof of the Cauchy criterion. ♣

## Exercises

1. Show that every Cauchy sequence is bounded.

**Solution:** Let  $(x_n)$  be a Cauchy sequence. Taking  $\varepsilon = 1$ , there is an integer  $N$  such that

$$n > N \Rightarrow |x_n - x_N| < 1.$$

Since  $|x_n| - |x_N| \leq |x_n - x_N|$ , we have

$$n > N \Rightarrow |x_n| < |x_N| + 1.$$

Thus  $|x_n|$  is bounded by  $\max\{|x_1|, |x_2|, \dots, |x_{N-1}|, |x_N| + 1\}$ . ♣

2. Let  $x_1 = 1$ ,  $x_2 = 2$ , and  $x_n = (x_{n-1} + x_{n-2})/2$  for all  $n \geq 3$ . Show that  $(x_n)$  is a Cauchy sequence.

**Solution:** Since for  $n \geq 3$ ,  $x_n - x_{n-1} = -(x_{n-1} - x_{n-2})/2$ , we use the induction on  $n$  to obtain  $x_n - x_{n+1} = (-1)^n / 2^{n-1}$  for all  $n \in \mathbf{N}$ . Hence, if  $m > n$ , then

$$\begin{aligned} |x_n - x_m| &\leq |x_n - x_{n+1}| + |x_{n+1} - x_{n+2}| + \dots + |x_{m-1} - x_m| \\ &= \sum_{k=n}^{m-1} \frac{1}{2^{k-1}} = \frac{1}{2^{n-1}} \sum_{k=0}^{m-n-1} \frac{1}{2^k} \\ &= \frac{1}{2^{n-1}} \frac{1 - (1/2)^{m-n}}{1 - (1/2)} < \frac{1}{2^{n-1}} \frac{1}{1 - (1/2)} = \frac{1}{2^{n-2}}. \end{aligned}$$

Since  $1/2^{n-2}$  decreases monotonically with  $n$ , it is possible to choose  $N$  for any  $\varepsilon > 0$  such that  $(1/2^{N-2}) < \varepsilon$ . We thus conclude that

$$m > n \geq N \Rightarrow |x_n - x_m| < \left(\frac{1}{2}\right)^{n-2} < \left(\frac{1}{2}\right)^{N-2} < \varepsilon,$$

which means that  $(x_n)$  is a Cauchy sequence. ♣

3. Suppose that the two sequences  $(x_n)$  and  $(y_n)$  converge to a common limit  $c$  and consider their **shuffled sequence**  $(z_n)$  defined by

$$(z_1, z_2, z_3, z_4, \dots) = (x_1, y_1, x_2, y_2, \dots).$$

Show that the sequence  $(z_n)$  also converges to  $c$ .

**Solution:** Let  $\varepsilon$  be any positive number. Since  $x_n \rightarrow c$  and  $y_n \rightarrow c$ , there are two positive integers  $N_1$  and  $N_2$  such that

$$n \geq N_1 \Rightarrow |x_n - c| < \varepsilon \text{ and } n \geq N_2 \Rightarrow |y_n - c| < \varepsilon.$$

Define  $N = \max\{N_1, N_2\}$ . Since  $x_k = z_{2k-1}$  and  $y_k = z_{2k}$  for all  $k \in \mathbf{N}$ , we have

$$k \geq N \Rightarrow |x_k - c| = |z_{2k-1} - c| < \varepsilon \text{ and } |y_k - c| = |z_{2k} - c| < \varepsilon.$$

Hence,  $n \geq 2N - 1 \Rightarrow |z_n - c| < \varepsilon$ , which just means  $\lim z_n = c$ . ♣

4. Show that  $\lim_{n \rightarrow \infty} (a^n/n^k) \rightarrow \infty$ , where  $a > 1$  and  $k > 0$ .

**Solution:** We consider three cases in turn: (i)  $k = 1$ , (ii)  $k < 1$ , and (iii)  $k > 1$ .

(i) Let  $k = 1$ . Then set  $a = 1 + h$  to obtain

$$a^n = (1 + h)^n = 1 + nh + \frac{n(n-1)}{2}h^2 + \cdots > \frac{n(n-1)}{2}h^2,$$

which results in

$$a^n/n = (1 + h)^n/n > (n-1)h^2/2 \rightarrow \infty. \quad (n \rightarrow \infty).$$

(ii) The case of  $k < 1$  is trivial since  $a^n/n^k > a^n/n$  for any  $n > 1$ .

(iii) If  $k > 1$ , then  $a^{1/k} > 1$  since  $a > 1$ . Hence, it follows from the result of (i) that for any  $M > 1$ , we can find an  $n$  so that  $n > M \Rightarrow a^{1/k}/n > M$ . This means that

$$\frac{a^n}{n^k} = \left[ \frac{(a^{1/k})^n}{n} \right]^k > M^k > M,$$

which implies that  $a^n/n^k \rightarrow \infty$ . ♣

5. Let  $x_n = a^n/n!$  with  $a > 0$ . Show that the sequence  $(x_n)$  converges to 0.

**Solution:** Let  $k$  be a positive integer such that  $k > 2a$ , and define  $c = a^k/k!$ . Then for any  $a > 0$  and for any  $n > k$ , we have

$$\frac{a^n}{n!} = c \frac{a}{k+1} \cdot \frac{a}{k+2} \cdots \frac{a}{n} < \frac{c}{2^{n-k}} = \frac{c \cdot 2^k}{2^n} < \frac{c \cdot 2^k}{n}. \quad (2.8)$$

Since (2.8) holds for a sufficiently large  $n$  ( $> k$ ), it also holds for  $n$  satisfying  $n > 2^k c/\varepsilon$ , where  $\varepsilon$  is an arbitrarily small number. In the latter case, we have

$$\frac{a^n}{n!} < \frac{2^k c}{n} < \varepsilon,$$

which means that

$$\lim_{n \rightarrow \infty} x_n = \lim_{n \rightarrow \infty} \frac{a^n}{n!} = 0. \quad \clubsuit$$

## 2.3 Infinite Series of Real Numbers

### 2.3.1 Limits of Infinite Series

This section focuses on convergence properties of **infinite series**. The importance of this issue will become apparent, particularly in connection with

certain branches of functional analysis such as Hilbert space theory and orthogonal polynomial expansions, where infinite series of numbers (or of functions) enter quite often (see Chaps. 4 and 5).

To begin with, we briefly review the basic properties of infinite series of real numbers. Assume an infinite sequence  $(a_1, a_2, \dots, a_n, \dots)$  of real numbers. We can then form another infinite sequence  $(A_1, A_2, \dots, A_n, \dots)$  with the definition

$$A_n = \sum_{k=1}^n a_k.$$

Here,  $A_n$  is called the  $n$ th **partial sum** of the sequence  $(a_n)$ , and the corresponding infinite sequence  $(A_n)$  is called the **sequence of partial sums** of  $(a_n)$ . The infinite sequence  $(A_n)$  may or may not be convergent, which depends on the features of  $(a_n)$ .

Let us introduce an **infinite series** defined by

$$\sum_{k=1}^{\infty} a_k = a_1 + a_2 + \dots \quad (2.9)$$

The infinite series (2.9) is said to converge if and only if the sequence  $(A_n)$  converges to the limit denoted by  $A$ . In other words, the series (2.9) converges if and only if the **sequence of the remainder**  $R_{n+1} = A - A_n$  converges to zero. When  $(A_n)$  is convergent, its limit  $A$  is called the **sum of the infinite series** of (2.9), and we may write

$$\sum_{k=1}^{\infty} a_k = \lim_{n \rightarrow \infty} \sum_{k=1}^n a_k = \lim_{n \rightarrow \infty} A_n = A.$$

Otherwise, the series (2.9) is said to diverge.

The limit of the sequence  $(A_n)$  is formally defined in line with Cauchy's procedure as shown below.

♠ **Limit of a sequence of partial sums:**

The sequence of partial sums  $(A_n)$  has a limit  $A$  if for any small  $\varepsilon > 0$ , there exists a number  $N$  such that

$$n > N \Rightarrow |A_n - A| < \varepsilon. \quad (2.10)$$

*Examples* 1. The infinite series  $\sum_{k=1}^{\infty} \left( \frac{1}{k} - \frac{1}{k+1} \right)$  converges to 1 because

$$A_n = \sum_{k=1}^n \left( \frac{1}{k} - \frac{1}{k+1} \right) = 1 - \frac{1}{n+1} \rightarrow 1 \quad (n \rightarrow \infty).$$

2. The series  $\sum_{k=1}^{\infty} (-1)^k$  diverges because the sequence

$$A_n = \sum_{k=1}^n (-1)^k = \begin{cases} 0 & n \text{ (is even)}, \\ -1 & n \text{ (is odd)} \end{cases}$$

does not approach any limit.

3. The series  $\sum_{k=1}^{\infty} 1 = 1 + 1 + 1 + \cdots$  diverges since the sequence  $A_n = \sum_{k=1}^n 1 = n$  increases without limit as  $n \rightarrow \infty$ .

### 2.3.2 Cauchy Criterion for Infinite Series

The following is a direct application of the Cauchy criterion to the sequence  $(A_n)$ , which consists of the partial sum  $A_n = \sum_{k=1}^n a_k$ :

♠ **Cauchy criterion for infinite series:**

The sequence of partial sums  $(A_n)$  converges if and only if for any small  $\varepsilon > 0$  there exists a number  $N$  such that

$$n, m > N \Rightarrow |A_n - A_m| < \varepsilon. \quad (2.11)$$

Similarly to the case of real sequences, the Cauchy criterion alluded to above provides a necessary and sufficient condition for convergence of the sequence  $(A_n)$ . Moreover, from the definition, it also gives a necessary and sufficient condition for convergence of an infinite series  $\sum_{k=1}^{\infty} a_k$ . Below is an important theorem associated with the latter statement.

♠ **Theorem:**

If an infinite series  $\sum_{k=1}^{\infty} a_k$  is convergent, then

$$\lim_{n \rightarrow \infty} a_n = 0.$$

*Proof* From hypothesis, we have

$$\lim_{n \rightarrow \infty} \sum_{k=1}^n a_k = \lim_{n \rightarrow \infty} A_n = A.$$

Hence,

$$\lim_{n \rightarrow \infty} a_n = \lim_{n \rightarrow \infty} (A_n - A_{n-1}) = A - A = 0. \quad \clubsuit$$

According to the theorem above,  $\lim a_n = 0$  is a necessary condition for the convergence of  $A_n$ . However, it is not a sufficient condition, as shown in the following example.

*Examples* Let  $a_k = 1/\sqrt{k}$ . Although  $\lim_{k \rightarrow \infty} a_k = 0$ , the corresponding infinite series  $\sum a_k$  diverges, as seen from

$$\begin{aligned} \sum_{k=1}^n a_k &= 1 + \frac{1}{\sqrt{2}} + \cdots + \frac{1}{\sqrt{n}} \\ &\geq \frac{1}{\sqrt{n}} + \frac{1}{\sqrt{n}} + \cdots + \frac{1}{\sqrt{n}} = \frac{n}{\sqrt{n}} = \sqrt{n} \rightarrow \infty. \end{aligned}$$

**Remark.** The contraposition of the previous theorem serves as a **divergent test** of the infinite series in question; we can say that

$$\lim_{n \rightarrow \infty} a_n \neq 0 \Rightarrow \sum_{k=1}^{\infty} a_k \text{ is divergent.}$$

### 2.3.3 Absolute and Conditional Convergence

Assume an infinite series

$$\sum_{k=1}^{\infty} a_k, \tag{2.12}$$

and an associated auxiliary series

$$\sum_{k=1}^{\infty} |a_k|, \tag{2.13}$$

in the latter of which all terms are positive. If the series (2.13) converges, then the series (2.12) is said to **converge absolutely**. The necessary and sufficient condition for absolute convergence of (2.12) is obtained by replacing  $A_n$  in (2.11) by

$$B_n = |a_1| + |a_2| + \cdots + |a_n|.$$

If the series (2.13) diverges and the original series (2.12) converges, we say that the series (2.12) **converges conditionally**. These results are summarized by the statement below.

♠ **Absolute convergence:**

The infinite series  $\sum a_k$  is absolutely convergent if  $\sum |a_k|$  is convergent.

♠ **Conditional convergence:**

The infinite series  $\sum a_k$  is conditionally convergent if  $\sum a_k$  is convergent and  $\sum |a_k|$  is divergent.

*Examples* The infinite series

$$\sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} \quad (2.14)$$

converges conditionally, since it converges while its absolute-value series  $\sum_{k=1}^{\infty} |(-1)^{k+1}/k| = \sum_{k=1}^{\infty} (1/k)$  diverges. See Exercises **1** and **2** in this section.

The following is an important theorem that we use many times in the remainder of this book.

♠ **Theorem:**

An infinite series converges if it converges absolutely.

*Proof* Suppose that the series  $(B_n)$  consisting of

$$B_n = \sum_{k=1}^n |a_k|$$

converges as  $n \rightarrow \infty$ . This means that for any  $\varepsilon > 0$  a number  $N$  exists such that

$$n, m > N \Rightarrow |B_n - B_m| < \varepsilon. \quad (2.15)$$

Assuming  $n \geq m$ , we rewrite the left-hand inequality in (2.15) as

$$\begin{aligned} |B_n - B_m| &= |a_{m+1}| + |a_{m+2}| + \cdots + |a_n| \\ &\geq |a_{m+1} + a_{m+2} + \cdots + a_n| \\ &= |A_n - A_m|, \end{aligned} \quad (2.16)$$

where we used the law of inequalities for sums. Hence, it follows from (2.15) and (2.16) that

$$n, m > N \Rightarrow |A_n - A_m| < \varepsilon,$$

which means that the series  $\sum a_k$  converges. ♣

The converse of the above theorem is not true. Below we present a well-known example of a convergent series that is not absolutely convergent.

### 2.3.4 Rearrangements

Observe that the conditionally convergent series (2.14) expressed by

$$1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \cdots, \quad (2.17)$$

may be rearranged in a number of ways, such as

$$1 + \frac{1}{3} - \frac{1}{2} + \frac{1}{5} - \frac{1}{4} + \cdots \quad (2.18)$$

or

$$-\frac{1}{2} - \frac{1}{8} + \frac{1}{7} + 1 + \frac{1}{3} - \cdots \quad (2.19)$$

or in any other way in which the terms  $1, -1/2, 1/3, -1/4, \dots$  are added in a certain order. Series such as (2.18) and (2.19) are called **rearrangements** of the series (2.17).

Of importance is the fact that rearranging procedures may change the convergence property of a conditionally convergent series; in what way this happens depends on the nature of the original series, as we shall now see. Suppose a series  $\sum a_n$  to be conditionally convergent. Then, the sum of its positive terms or that of its negative term goes, respectively to  $+\infty$  or  $-\infty$ ; otherwise the original series would diverge or converge absolutely. Let  $(b_n)$  and  $(c_n)$  be, respectively, the subsequences of positive and negative terms of  $(a_n)$ . Since  $\sum_{k=1}^n b_k$  is monotonically increasing with respect to  $n$ , there is a positive integer  $m_1$  such that

$$\sum_{k=1}^{m_1} b_k \geq 1 - c_1.$$

Here the right-hand side is positive since  $c_1$  is negative. We rewrite it as

$$\sum_{k=1}^{m_1} b_k + c_1 \geq 1.$$

Similarly, there is an integer  $m_2 > m_1$  such that

$$\sum_{k=1}^{m_2} b_k + c_2 \geq 1.$$

Continue on the same process for  $m_3, m_4, \dots, m_n$  and take the sum of each side to obtain

$$\sum_{k=1}^{m_n} b_k + \sum_{k=1}^n c_k \geq n. \quad (2.20)$$

Note that the left-hand side is a partial sum of the rearrangement of the sequence  $(a_k)$  that may, for instance, take the form of

$$(b_1, b_2, \dots, b_{m_1}, c_1, b_{m_1+1}, b_{m_1+2}, \dots, b_{m_2}, c_2, \dots). \quad (2.21)$$

Clearly, the left-hand side of (2.20) diverges as  $n \rightarrow \infty$ , which means that the rearrangement (2.21) diverges. Therefore, the conditionally convergent series may become divergent through the rearranging procedure. In fact, the discussion above serves as part of the proof of the theorem below.

♠ **Riemann theorem:**

Given any conditionally convergent series and any  $r \in \overline{\mathbf{R}} = \mathbf{R} \cup \infty$ , there is a rearrangement of the series that converges to  $r$ .

*Proof* The case of  $r = \infty$  was proved in the previous discussion. Now let  $r \in \mathbf{R}$  and assume that  $(b_n)$  and  $(c_n)$  is the subsequence of positive and negative terms, respectively, in the same order in which they appear in  $(a_n)$ . It is possible to obtain the smallest sum such that

$$s_1 = \sum_{k=1}^{m_1} b_k$$

exceeds  $r$ . Then, add the least number of negative terms  $c_k$  to obtain the largest sum. Such that

$$s_2 = \sum_{k=1}^{m_1} b_k + \sum_{k=1}^{n_1} c_k$$

is less than  $r$ . Proceeding in this fashion, we obtain a sequence  $s_1, s_2, s_3, \dots$  that converges to  $r$ , since

$$\lim_{n \rightarrow \infty} b_n = \lim_{n \rightarrow \infty} c_n = 0.$$

This result is the case for an arbitrary real number  $r$ . Hence, the proof is complete. ♣

## Exercises

1. Determine the convergent property of the series

$$\sum_{k=1}^{\infty} \frac{1}{k}. \quad (2.22)$$

This is known as a **harmonic series**.

**Solution:** Let  $A_n = \sum_{k=1}^n (1/k)$ . We then have

$$A_{2n} - A_n = \frac{1}{n+1} + \frac{1}{n+2} + \cdots + \frac{1}{2n} \geq \frac{1}{2n} \times n = \frac{1}{2},$$

which implies that the sequence  $(A_n)$  is not a Cauchy sequence. Thus, view of the Cauchy criterion, the harmonic series (2.22) diverges. ♣

**2.** Determine the convergence of the series

$$\sum_{k=1}^{\infty} \frac{1}{k^p}. \quad (2.23)$$

This is called a **hyperharmonic series** (or **zeta function**) and is denoted by  $\xi(p)$ .

**Solution:** When  $p \leq 1$ , a partial sum  $A_{2^n}$  consisting of the first  $2^n$  terms reads

$$\begin{aligned} A_{2^n} &= \left(1 + \frac{1}{2^p}\right) + \left(\frac{1}{3^p} + \frac{1}{4^p}\right) + \left(\frac{1}{5^p} + \cdots + \frac{1}{8^p}\right) + \cdots \\ &\quad + \left[\frac{1}{(2^{n-1}+1)^p} + \cdots + \frac{1}{(2^n)^p}\right] \\ &\geq \left(1 + \frac{1}{2}\right) + \left(\frac{1}{3} + \frac{1}{4}\right) + \left(\frac{1}{5} + \cdots + \frac{1}{8}\right) + \cdots \\ &\quad + \left[\frac{1}{(2^{n-1}+1)} + \cdots + \frac{1}{2^n}\right] \\ &\geq \frac{1}{2} + \frac{1}{4} \times 2 + \frac{1}{8} \times 4 + \cdots + \frac{1}{2^n} \times 2^{n-1} = \frac{n}{2}. \end{aligned}$$

This means that the series (2.23) diverges for  $p \leq 1$ .

For  $p > 1$ , we have

$$\begin{aligned} A_{2^{n+1}-1} &= 1 + \left(\frac{1}{2^p} + \frac{1}{3^p}\right) + \left(\frac{1}{4^p} + \cdots + \frac{1}{7^p}\right) + \cdots \\ &\quad + \left[\frac{1}{(2^n)^p} + \cdots + \frac{1}{(2^{n+1}-1)^p}\right] \\ &< 1 + \frac{1}{2^p} \times 2 + \frac{1}{4^p} \times 4 + \cdots + \frac{1}{(2^n)^p} \times 2^n \\ &< \sum_{k=0}^n \left(\frac{1}{2^{p-1}}\right)^k = \frac{1 - (1/2^{p-1})^{n+1}}{1 - (1/2^{p-1})} < \frac{2^{p-1}}{2^{p-1} - 1}. \end{aligned}$$

Hence, the monotonically increasing sequence  $\{A_n\}$  is bounded above and is thus convergent. ♣

3. Determine the convergence of the series

$$\sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k}. \quad (2.24)$$

**Solution:** Let  $n$  be an even integer, say  $n = 2m$ . Then, it follows that

$$\begin{aligned} A_{2m} &= \sum_{k=1}^{2m} \frac{(-1)^{k+1}}{k} = \left(1 - \frac{1}{2}\right) + \left(\frac{1}{3} - \frac{1}{4}\right) + \cdots + \left(\frac{1}{2m-1} - \frac{1}{2m}\right) \\ &= \frac{1}{2 \cdot 1} + \frac{1}{4 \cdot 3} + \cdots + \frac{1}{2m(2m-1)}, \end{aligned}$$

which means that  $(A_{2m})$  is increasing with  $m$ . In addition, we have

$$A_{2m} = 1 - \left(\frac{1}{2} - \frac{1}{3}\right) - \left(\frac{1}{4} - \frac{1}{5}\right) - \cdots - \frac{1}{2m} < 1,$$

which indicates that  $(A_{2m})$  is bounded above. Hence,  $(A_{2m})$  converges to a limit  $A$ . Further more, since  $A_{2m+1} = A_{2m} + 1/(2m+1)$ , the same discussion as above tells us that the sequence  $(A_{2m+1})$  also converges to the common limit  $A$ . By applying the result from shuffled sequences (see Exercise 3 in Sect. 2.1.2), we find that  $\lim A_n$  exists, so the series (2.24) converges. It is thus proven that the series converges conditionally. ♣

4. Suppose that the infinite series  $\sum_k a_k$  and  $\sum_k b_k$  are both convergent absolutely. Let  $(a_i b_j)$  be an infinite sequence in which the terms  $a_i b_j$  are arranged in an arbitrary order, say, as

$$(a_2 b_1, a_1 b_3, a_2 b_4, a_5 b_1, \dots).$$

Show that the sequence of the partial sums of  $(a_i b_j)$  converges absolutely regardless of the order of the terms  $a_i b_j$ .

**Solution:** Let  $m$  and  $n$  be the maximum values of  $i$  and  $j$ , respectively, that are involved in the partial sum  $\sum_{(i,j)} a_i b_j$ ; here  $(i, j)$  denotes the possible combinations of  $i$  and  $j$  that are arranged in the same order as in the sequence  $(a_i b_j)$ . The partial

sum is a portion of the product of the finite sums given by  $(\sum_{i=1}^m a_i) \left( \sum_{j=1}^n b_j \right)$ . Hence, we have

$$\left| \sum_{(i,j)} a_i b_j \right| = \sum_{(i,j)} |a_i b_j| = \left| \sum_i^m a_i \right| \cdot \left| \sum_j^n b_j \right| \leq \sum_{i=1}^m |a_i| \sum_{j=1}^n |b_j|. \quad (2.25)$$

From hypothesis, the left-hand side in (2.25) converges as  $m, n \rightarrow \infty$ . This means that the partial sum  $\sum_{(i,j)} |a_i b_j|$  is bounded above.

In addition, it is obviously increasing. Therefore,  $\sum_{(i,j)} |a_i b_j|$  converges (i.e.,  $\sum_{(i,j)} a_i b_j$  converges absolutely) independently of the order of  $i$  and  $j$  in the sequence of  $(a_i b_j)$ . ♣

5. Show that rearrangements of absolutely convergent series always converge absolutely to the same limit.

**Solution:** Let  $\sum_{k=1}^{\infty} a_k$  be absolutely convergent and assume that  $\sum_{k=1}^{\infty} b_k$  is its rearrangement. Define  $A_n = \sum_{k=1}^n |a_k|$ ,  $A = \lim_{n \rightarrow \infty} A_n$ ,  $B_n = \sum_{k=1}^n |b_k|$ , and let  $\varepsilon > 0$ . By hypothesis, there is an integer  $N$  such that  $|A - A_N| = |a_{N+1}| + |a_{N+2}| + \cdots < \frac{\varepsilon}{2}$ . Now we choose the integer  $M$  so that all the terms  $a_1, a_2, \dots, a_N$  appear in the first  $M$  terms of the rearranged series, i.e., within the finite sequence  $(b_1, b_2, \dots, b_M)$ . Hence, these terms do not contribute to the difference  $B_m - A_N$ , where  $m \geq N$ . Consequently, we obtain

$$\begin{aligned} m \geq N &\Rightarrow |B_m - A_N| \leq |a_{N+1}| + |a_{N+2}| + \cdots < \frac{\varepsilon}{2} \\ &\Rightarrow |A - B_m| \leq |A - A_N| + |A_N - B_m| < \varepsilon, \end{aligned}$$

which shows that  $\lim_{n \rightarrow \infty} B_n = A$ . ♣

## 2.4 Convergence Tests for Infinite Real Series

### 2.4.1 Limit Tests

This section covers the important tests for convergence of infinite series. In general, these tests provide sufficient, not necessary, conditions for convergence. This is in contrast to the Cauchy criterion, which provides a necessary and sufficient condition for convergence, though it is difficult to apply in practice. The first test to be shown is called the **limit test**, by which we can examine the absolute convergence of infinite series quite easily.

♠ **Limit test for convergence:**

If

$$\lim_{k \rightarrow \infty} k^p a_k \text{ exists for some } p > 1,$$

then  $\sum_{k=1}^{\infty} a_k$  converges absolutely (and thus converges ordinary).

*Proof* By hypothesis, we set  $\lim_{k \rightarrow \infty} k^p a_k = A$  for certain  $p > 1$ , which implies that

$$\lim_{k \rightarrow \infty} k^p |a_k| = |A|.$$

Hence, there exists an integer  $m$  such that

$$k > m \Rightarrow k^p |a_k| - |A| < 1,$$

or equivalently,

$$k > m \Rightarrow |a_k| < \frac{|A| + 1}{k^p}. \quad (2.26)$$

We know that the series  $\sum_{k=m}^{\infty} 1/k^p$  converges for any  $p > 1$  (see Exercise 2 in Sect. 2.3). Thus it follows from (2.26) that the series  $\sum_{k=m}^{\infty} |a_k|$  also converges, from which the desired conclusion follows at once. ♣

There is a counterpart of the limit test for convergence that determines *divergence* properties of series as follows.

♠ **Limit test for divergence:**

If

$$\lim_{k \rightarrow \infty} k a_k \neq 0,$$

then  $\sum_{k=1}^{\infty} a_k$  diverges. The test fails if the limit equals zero.

*Proof* Suppose  $\lim k a_k = A > 0$ . Then there exists an integer  $m$  such that

$$k \geq m \Rightarrow k a_k > \frac{A}{2}.$$

Hence, by employing the result from harmonic series (see Exercise 1 in Sect. 2.3), we obtain

$$\sum_{k=m}^{\infty} a_k > \sum_{k=m}^{\infty} \frac{1}{k} = \infty,$$

from which the desired result follows. The same procedure can be applied to the case of  $A < 0$ , in which case the series  $\sum_{k=1}^{\infty} (-a_k)$  may be treated by the procedure above. The proof is thus complete. ♣

*Remark.*

1. The test is valid even when  $A$  goes to infinity.
2. The divergence test described above is inconclusive when  $\lim ka_k = 0$ . To see why, consider the two series

$$\sum_{k=1}^{\infty} \frac{1}{k^2} \quad \text{and} \quad \sum_{k=2}^{\infty} \frac{1}{k \log k}.$$

The former converges and the latter diverges, but both yield  $\lim ka_k = 0$ .

### 2.4.2 Ratio Tests

The following provides another test for absolute convergence of infinite series that is sometimes easier to use than the previous one.

♠ **Ratio test:**

A series  $\sum_{k=0}^{\infty} a_k$  converges absolutely (and thus converges ordinary) if

$$\limsup_{k \rightarrow \infty} \left| \frac{a_{k+1}}{a_k} \right| < 1 \quad (2.27)$$

and diverges if

$$\limsup_{k \rightarrow \infty} \left| \frac{a_{k+1}}{a_k} \right| > 1. \quad (2.28)$$

If the limit superior is 1, the test is inconclusive.

*Remark.* When  $|a_{k+1}/a_k|$  converges, the limits superior used in (2.27) and (2.28) reduce to the ordinary limits.

*Proof (i)* Suppose that  $\ell \equiv \limsup_{k \rightarrow \infty} \left| \frac{a_{k+1}}{a_k} \right| < 1$ . Then, for any  $r \in (\ell, 1)$ , we can find the number  $m$  such that

$$k > m \Rightarrow \left| \frac{a_{k+1}}{a_k} \right| < r.$$

It follows that  $\left| \frac{a_{m+1}}{a_m} \right| \times \left| \frac{a_{m+2}}{a_{m+1}} \right| \times \cdots \times \left| \frac{a_{m+p}}{a_{m+p-1}} \right| < r^p$  or equivalently,

$|a_{m+p}| < r^p |a_m|$ , which holds for any  $p \in \mathbf{N}$ . Hence, we have

$$\sum_{p=1}^{\infty} |a_{m+p}| = \sum_{k=m+1}^{\infty} |a_k| < \sum_{p=1}^{\infty} r^p |a_m| = \frac{r}{1-r} |a_m|.$$

The last term is a finite constant. Therefore, the series  $\sum_{k=m+1}^{\infty} |a_k|$  remains finite and the series  $\sum_{k=0}^{\infty} a_k$  converges absolutely.

(ii) Next we assume that

$$\limsup_{k \rightarrow \infty} \left| \frac{a_{k+1}}{a_k} \right| = \ell > 1.$$

Then there is an integer  $m$  such that

$$k > m \Rightarrow \left| \frac{a_{k+1}}{a_k} \right| > 1.$$

That is,

$$k > m \Rightarrow |a_k| > |a_m| > 0,$$

which means that

$$\lim_{k \rightarrow \infty} a_k \neq 0.$$

In view of the remark in Sect. 2.3.2, the series  $\sum_{k=0}^{\infty} a_k$  diverges. ♣

### 2.4.3 Root Tests

We now give an alternative absolute-convergence test based on examining the  $k$ th root of  $|a_k|$ .

#### ♠ Root test:

A series  $\sum_{k=0}^{\infty} a_k$  converges absolutely (and ordinary) if

$$\limsup_{k \rightarrow \infty} \sqrt[k]{|a_k|} < 1$$

and diverges if

$$\limsup_{k \rightarrow \infty} \sqrt[k]{|a_k|} > 1.$$

If the limit superior is 1, the test fails and does not provide any information.

*Proof* Let  $r = \limsup_{k \rightarrow \infty} \sqrt[k]{|a_k|}$ . We first prove that the series converges absolutely if  $r < 1$ . We choose a positive number  $c \in (r, 1)$ . Then there is a positive integer  $N$  such that

$$k \geq N \Rightarrow \sqrt[k]{|a_k|} < c \Rightarrow |a_k| < c^k.$$

Since the geometric series  $\sum c^k$  with  $c < 1$  converges,  $\sum |a_k|$  converges, so that  $\sum a_k$  converges absolutely.

When  $r > 1$ , it follows from the definition of the limit superior (see Sect. 2.1.4) that there are an infinite number of terms of  $\sqrt[k]{|a_k|}$  greater than 1. This implies that  $\lim a_k \neq 0$ , which means that the infinite series  $\sum a_k$  diverges. ♣

*Examples* Assume the series

$$\sum_{k=0}^{\infty} a_k = 1 - \frac{1}{2} + \frac{1}{4^2} - \frac{1}{2^3} + \frac{1}{4^4} - \frac{1}{2^5} + \cdots. \quad (2.29)$$

Since

$$\sqrt[0]{|a_0|} = 1, \quad \sqrt[1]{|a_1|} = \frac{1}{2}, \quad \sqrt[2]{|a_2|} = \frac{1}{4}, \quad \sqrt[3]{|a_3|} = \frac{1}{2}, \quad \sqrt[4]{|a_4|} = \frac{1}{4}, \dots,$$

we have

$$\limsup_{k \rightarrow \infty} \sqrt[k]{a_k} = \frac{1}{2} < 1.$$

Thus the series (2.29) converges (absolutely and ordinary).

#### 2.4.4 Alternating Series Test

All the convergence tests presented so far are tests for absolute convergence, which assumes ordinary convergence. Nonetheless, certain kinds of series can exhibit conditional convergence, i.e., ordinary convergence with absolute divergence, whose convergence properties cannot be addressed by the tests given thus far. Hence, the significance of the test described below, known as the **alternating series test**, is that it may be used to test the conditional convergence of some absolutely divergent series.

We say that  $(x_k)$  is an **alternating sequence** if the sign of  $x_k$  is different from that of  $x_{k+1}$  for every  $k$ . The resulting series  $\sum x_k$  is called the **alternating series**, whose convergence properties are partly determined by the following theorem:

##### ♠ Alternating series test:

An alternating series given by

$$a_1 - a_2 + a_3 - a_4 + \cdots = \sum_{k=1}^{\infty} (-1)^{k+1} a_k \quad \text{with } a_k > 0 \text{ for all } k$$

converges if

$$a_k > a_{k+1} \quad \text{and} \quad \lim_{k \rightarrow \infty} a_k = 0.$$

*Proof* First we show that the sequence of partial sums  $S_n$  converges. It follows that

$$A_{2n} = (a_1 - a_2) + (a_3 - a_4) + \cdots + (a_{2n-1} - a_{2n}).$$

Since  $a_k - a_{k+1} > 0$  for all  $k$ , the sequence  $A_{2n}$  is increasing. It is also bounded above because

$$A_{2n} = a_1 - (a_2 - a_3) - (a_4 - a_5) - \cdots - (a_{2n-2} - a_{2n-1}) - a_{2n} \leq a_1 \quad (2.30)$$

for all  $n \in \mathbf{N}$ . Thus,  $\lim A_{2n}$  exists and we call it  $A$ . On the other hand, we have

$$|A_{2n+1} - A| = |A_{2n}a_{2n+1} - A| \leq |A_{2n} - A| + |a_{2n+1}|.$$

In the limit as  $n \rightarrow \infty$ , the left-hand side vanishes so that we obtain  $\lim A_{2n+1} = A$ . Therefore, we conclude that  $S_n \rightarrow S$  ♣

## Exercises

1. Show that  $\sum_{k=1}^{\infty} \frac{(k+1)^{1/2}}{(k^5 + k^3 - 1)^{1/3}}$  converges.

**Solution:** Taking  $p = 7/6 > 1$  into the limit test for convergence, we have

$$\lim_{k \rightarrow \infty} k^{7/6} a_k = \lim_{k \rightarrow \infty} \frac{(1 + k^{-1})^{1/2}}{(1 - k^{-2} + k^{-5})^{1/3}} = 1. \quad \clubsuit$$

2. Show that  $\sum_{k=1}^{\infty} (-1)^k \frac{\log k}{k^2}$  converges.

**Solution:** With use of the limit test for convergence by taking  $p = 3/2$ , we obtain

$$\lim_{k \rightarrow \infty} k^{3/2} a_k = \lim_{k \rightarrow \infty} (-1)^k \frac{\log k}{\sqrt{k}} = 0. \quad \clubsuit$$

3. Show that  $\sum_{k=1}^{\infty} \frac{k \log k}{1 + k^2}$  diverges.

**Solution:** From the limit test for divergence, we have

$$\lim_{k \rightarrow \infty} k a_k = \lim_{k \rightarrow \infty} \frac{k^2 \log k}{1 + k^2} = \infty. \quad \clubsuit$$

4. Show that  $\sum_{k=0}^{\infty} \frac{(k!)^2}{(2k)!}$  converges.

**Solution:** The ratio test yields

$$\begin{aligned} \left| \frac{a_{k+1}}{a_k} \right| &= \frac{(2k)! [(k+1)!]^2}{(k!)^2 (2k+2)!} = \frac{(k+1)^2}{(2k+2)(2k+1)} \\ &= \frac{(1 + \frac{1}{k})^2}{(2 + \frac{2}{k})(2 + \frac{1}{k})} \rightarrow \frac{1}{4} < 1 \quad (k \rightarrow \infty). \quad \clubsuit \end{aligned}$$

5. Show that  $\sum_{k=1}^{\infty} \left(1 + \frac{1}{k}\right)^{-k^2}$  converges.

**Solution:** The root test yields

$$\left[ \left(1 + \frac{1}{k}\right)^{-k^2} \right]^{1/k} = \frac{1}{[1 + (1/k)]^k} \rightarrow \frac{1}{e} < 1.$$

## Real Functions

---

**Abstract** Infinite sequences and series of real functions are encountered frequently in mathematical physics. The convergence of such sequences and series does not generally preserve the nature of their constituents; e.g., a sequence of “continuous” functions can converge into a “discontinuous” function. In this chapter, we show that this is not true in cases of uniform convergence (Sect. 3.2.2), which is a special class of convergence that preserves the continuity, integrability, and differentiability of the constituent functions of sequences and series, as we explain in detail in Sects. 3.2.4–3.2.6.

### 3.1 Fundamental Properties

#### 3.1.1 Limit of a Function

Having discussed the limits of sequences and series of real numbers, we now turn our attention to the limit of functions. Let  $A$  be a real number and  $f(x)$  a real-valued function of a real variable  $x \in \mathbf{R}$ . A formal notation of the above function is given by the mapping relation  $f : \mathbf{R} \rightarrow \mathbf{R}$ . The statement “*the limit of  $f(x)$  at  $x = a$  is  $A$* ” means that the value of  $f(x)$  can be set as close to  $A$  as desired by setting  $x$  sufficiently close to  $a$ . This is stated formally by the following definition.

♠ **The limit of a function:**

A function  $f(x)$  is said to have the **limit**  $A$  as  $x \rightarrow a$  if and only if for every  $\varepsilon > 0$ , there exists a number  $\delta > 0$  such that

$$|x - a| < \delta \Rightarrow |f(x) - A| < \varepsilon. \quad (3.1)$$

The limit of  $f(x)$  is written symbolically as

$$\lim_{x \rightarrow a} f(x) = A$$

or

$$f(x) \rightarrow A \text{ for } x \rightarrow a.$$

If the first inequality in (3.1) is replaced by  $0 < x - a < \delta$  (or  $0 < a - x < \delta$ ), we say that  $f(x)$  approaches  $A$  as  $x \rightarrow a$  from above (or below) and write

$$\lim_{x \rightarrow a+} f(x) = A \quad \left( \text{or } \lim_{x \rightarrow a-} f(x) = A \right).$$

This is called the **right-hand** (or **left-hand**) **limit** of  $f(x)$ . The two together are known as **one-sided limits**.

A necessary and sufficient condition for the existence of  $\lim_{x \rightarrow a} f(x)$  is shown below.

♠ **Theorem:**

The limit of  $f(x)$  at  $x = a$  exists if and only if

$$\lim_{x \rightarrow a+} f(x) = \lim_{x \rightarrow a-} f(x). \quad (3.2)$$

*Proof* If  $\lim_{x \rightarrow a} f(x)$  exists and is equal to  $A$ , it readily follows that

$$\lim_{x \rightarrow a+} f(x) = \lim_{x \rightarrow a-} f(x) = A. \quad (3.3)$$

We now consider the converse. Assume that (3.2) holds. This obviously means that both one-sided limits exist at  $x = a$ . Hence, given  $\varepsilon > 0$ , we have  $\delta_1 > 0$  and  $\delta_2 > 0$  such that

$$\begin{aligned} 0 < x - a < \delta_1 &\Rightarrow |f(x) - A| < \varepsilon, \\ 0 < a - x < \delta_2 &\Rightarrow |f(x) - A| < \varepsilon. \end{aligned}$$

Let  $\delta = \min\{\delta_1, \delta_2\}$ . If  $x$  satisfies  $0 < |x - a| < \delta$ , then either

$$0 < x - a < \delta \leq \delta_1 \text{ or } 0 < a - x < \delta \leq \delta_2.$$

In either case, we have  $|f(x) - A| < \varepsilon$ . That is, we have seen that for a given  $\varepsilon$ , there exists  $\delta$  such that

$$0 < |x - a| < \delta \Rightarrow |f(x) - A| < \varepsilon.$$

Therefore we conclude that

$$\text{Equation (3.2) holds} \Rightarrow \lim_{x \rightarrow a} f(x) = A,$$

and the proof is complete. ♣

### 3.1.2 Continuity of a Function

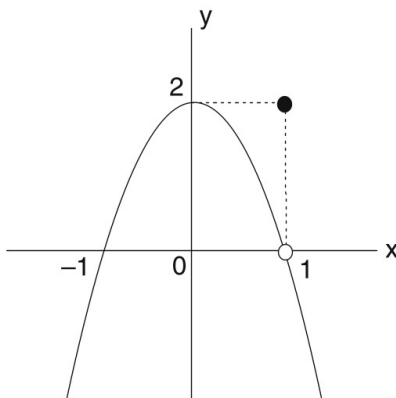
In general, the value of  $\lim_{x \rightarrow a} f(x)$  has nothing to do with the value (and the existence) of  $f(a)$ . For instance, the function given by

$$f(x) = \begin{cases} 2 - x^2 & x \neq 1, \\ 2 & x = 1 \end{cases}$$

gives

$$\lim_{x \rightarrow 1} f(x) = 0 \quad \text{and} \quad f(1) = 2,$$

which are quantitatively different from one another. This mismatch occurring at  $x = 1$  results in a lack of geographical continuity in the curve of  $y = f(x)$ , as depicted in Fig. 3.1. In mathematical language, continuity of the curve of  $y = f(x)$  is accounted for by the following statement.



**Fig. 3.1.** A discontinuous function  $y = f(x)$  at  $x = 1$

♠ **Continuous functions:**

The function  $f(x)$  is said to be **continuous** at  $x = a$  if and only if for every  $\varepsilon > 0$ , there exists  $\delta > 0$  such that

$$|x - a| < \delta \Rightarrow |f(x) - f(a)| < \varepsilon.$$

*Remark.* The definition noted above seems to be similar to the definition of the limit of  $f(x)$  at  $x = a$  (see Sect. 3.1.1). However, there is a crucial difference between them. When considering the limit of  $f(x)$  at  $x = a$ , we are only interested in the behavior of  $f(x)$  *in the vicinity of* the point  $a$ , not *just at*  $a$ . However, the continuity of  $f(x)$  at  $x = a$  requires the further condition that the value of  $f(x)$  just at  $x = a$  has to be defined. In symbols, we write

$$f(x) \text{ is continuous at } x = a \Rightarrow \lim_{x \rightarrow a-0} f(x) = \lim_{x \rightarrow a+0} f(x) = f(a).$$

We must emphasize that given a function  $f(x)$  on a domain  $D$ , the limit of  $f(x)$  is defined at limit points in  $D$  that may or may not lie in  $D$ . In contrast, the continuity of  $f(x)$  is defined only at points contained in  $D$ . An illustrative example is given below.

*Examples* Assume a function given by

$$f(x) = x \text{ for all but } x = 1.$$

It has a limit at  $x = 1$ ,

$$\lim_{x \rightarrow \infty} f(x) = 1,$$

but there is no way to examine its continuity because  $x = 1$  is out of the defining domain.

When  $f(x)$  is continuous, we can say that  $f(x)$  belongs to the class of functions designated by the symbol  $C$ . Then, it follows that

$$f(x) \in C \text{ at } x = a \iff \lim_{x \rightarrow a} f(x) = f(a).$$

If the symbol  $x \rightarrow a$  appearing in the right-hand statement is replaced by  $x \rightarrow a+$  (or  $x \rightarrow a-$ ),  $f(x)$  is said to be **continuous on the right** (or **left**) at  $x = a$ . We encounter the latter kind of a limit particularly when we consider the continuity of a function defined within a finite interval  $[a, b]$ ; we say that

$$\begin{aligned} f(x) \in C \text{ on } [a, b] &\iff \\ f(x) \in C \text{ on } (a, b) &\text{ and } \lim_{x \rightarrow a+} f(x) = f(a), \quad \lim_{x \rightarrow b-} f(x) = f(b). \end{aligned}$$

We also say that a function  $f(x)$  on  $[a, b]$  is **piecewise continuous** if

- (i)  $f(x)$  is continuous on  $[a, b]$  except at a finite number of points  $x_1, x_2, \dots, x_n$ ;
- (ii) at each of the points  $x_1, x_2, \dots, x_n$ , there exist both the left-hand and right-hand limits of  $f(x)$  defined by

$$f(x_k - 0) = \lim_{x \rightarrow x_k - 0} f(x), \quad f(x_k + 0) = \lim_{x \rightarrow x_k + 0} f(x).$$

### 3.1.3 Derivative of a Function

The following is a rigorous definition of the derivative of a real function.

#### ♠ Derivative of a function:

If the limit

$$\lim_{x \rightarrow a} \frac{f(x) - f(a)}{x - a}$$

exists, it is called the **derivative** of  $f(x)$  at  $x = a$  and is denoted by  $f'(a)$ . The function  $f(x)$  is said to be **differentiable** at  $x = a$  if  $f'(a)$  exists.

Similar to the case of one-sided limits, it is possible to define **one-sided derivatives** of real functions such as

$$f'(a+) = \lim_{x \rightarrow a+} \frac{f(x) - f(a)}{x - a},$$

$$f'(a-) = \lim_{x \rightarrow a-} \frac{f(x) - f(a)}{x - a}.$$

♠ **Theorem:**

If  $f(x)$  is differentiable at  $x = a$ , then it is continuous at  $x = a$ . (The converse is not true.)

*Proof* Assume  $x \neq a$ . Then

$$f(x) - f(a) = \frac{f(x) - f(a)}{x - a} (x - a).$$

From hypothesis, each function  $[f(x) - f(a)]/(x - a)$  as well as  $x - a$  has the limit at  $x = a$ . Hence, we obtain

$$\lim_{x \rightarrow a} [f(x) - f(a)] = \lim_{x \rightarrow a} \frac{f(x) - f(a)}{x - a} \cdot \lim_{x \rightarrow a} (x - a) = f'(a) \times 0 = 0.$$

Therefore,

$$\lim_{x \rightarrow a} f(x) = f(a),$$

i.e.,  $f(x)$  is continuous at  $x = a$ . That the converse is false can be seen by considering  $f(x) = |x|$ ; it is continuous at  $x = 0$  but not differentiable. ♣

The term  $C^n$  functions is used to indicate that all the derivatives on the order of  $\leq n$  exist; this is denoted by

$$f(x) \in C^n \iff f^{(n)}(x) \in C.$$

Such an  $f(x)$  is said to be a  $C^n$  function or to be of class  $C^n$ .

*Examples* 1.  $f(x) = \begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases}$

$$\Rightarrow f(x) \in C^0 (= C), \text{ but } f(x) \notin C^1 \text{ at } x = 0.$$

2.  $f(x) = \begin{cases} 0 & x < 0 \\ x^2 & x \geq 0 \end{cases}$

$$\Rightarrow f(x) \in C^1, \text{ but } f(x) \notin C^2 \text{ at } x = 0.$$

3. **Taylor series expansion** for functions  $f \in C^n$  is given by

$$f(x) = \sum_{k \leq n} \frac{1}{k!} \left. \frac{\partial f}{\partial x^k} \right|_{x=x_0} (x - x_0)^k + o(|x - x_0|^n).$$

### 3.1.4 Smooth Functions

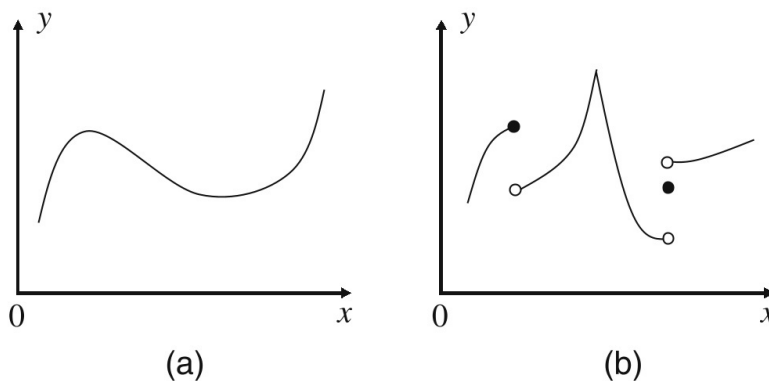
We now introduce a new class of functions for which the derivative is continuous over the defining domain.

♠ **Smooth functions:**

The function  $f(x)$  is said to be **smooth** for any  $x \in [a, b]$  if  $f'(x)$  exists and is continuous on  $[a, b]$ .

In geometrical language, the above statement means that the direction of the tangent changes continuously, without jumps, as it moves along the curve  $y = f(x)$  (see Fig. 3.2). Thus, the graph of a smooth function is a smooth curve without any point at which the curve has two distinct tangents.

Similar to the case of piecewise continuity, the function  $f(x)$  is said to be **piecewise smooth** on the interval  $[a, b]$  if  $f(x)$  and its derivatives are all piecewise continuous on  $[a, b]$ . The graph of a piecewise smooth function is either a continuous or a discontinuous curve; furthermore, it can have a finite number of points (called **corners**) at which the derivatives show jumps (see Fig. 3.2). Every piecewise smooth function  $f(x)$  is bounded and has a bounded derivative everywhere, except at its corners and points of discontinuity;  $f'(x)$  does not exist in the sense of continuity at any of these points.



**Fig. 3.2.** (a) A continuous function  $y = f(x)$ . (b) A piecewise smooth function  $y = f(x)$  having two discontinuous points and one corner

## 3.2 Sequences of Real Functions

### 3.2.1 Pointwise Convergence

In this section we focus on convergence properties of sequences consisting of real-valued *functions* of a real variable. Suppose that for each  $n \in \mathbf{N}$ , we have

a function  $f_n(x)$  defined on a domain  $D \subseteq \mathbf{R}$ . We then say that we have a sequence

$$(f_n(x) : n \in \mathbf{N})$$

of real-valued functions on  $D$ . If the sequence  $(f_n(x))$  converges for every  $x \in D$ , the sequence of functions is said to **converge pointwise** on  $D$ , and the function defined by

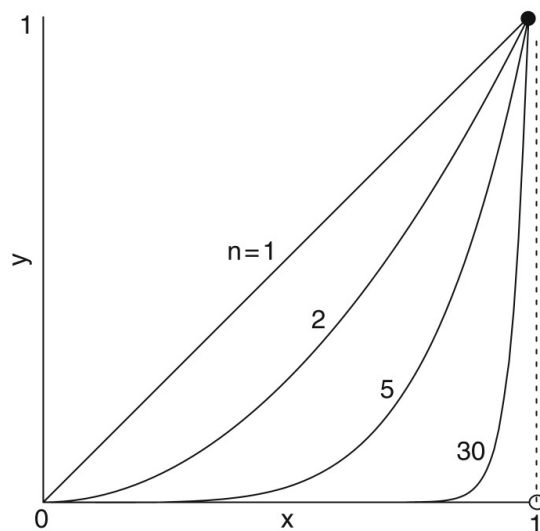
$$f(x) = \lim_{n \rightarrow \infty} f_n(x)$$

is called the **pointwise limit** of  $(f_n(x))$ . The formal definition is given below.

♠ **Pointwise convergence:**

The sequence of functions  $(f_n)$  is said to **converge pointwise** to  $f$  on  $D$  if, given  $\varepsilon > 0$ , there is a natural number  $N = N(\varepsilon, x)$  (which depends on  $\varepsilon$  and  $x$ ) such that

$$n > N \Rightarrow |f_n(x) - f(x)| < \varepsilon.$$



**Fig. 3.3.** Converging behavior of  $f_n(x) = x^n$  given in (3.4)

*Examples* Assume a sequence  $(f_n)$  consisting of the function

$$f_n(x) = x^n \tag{3.4}$$

that is defined on a closed interval  $[0, 1]$ . It follows that the sequence converges pointwise to

$$f(x) = \lim_{n \rightarrow \infty} f_n(x) = \begin{cases} 0 & \text{for } 0 \leq x < 1, \\ 1 & \text{at } x = 1. \end{cases} \quad (3.5)$$

See Fig. 3.3 for the converging behavior of  $f_n(x)$  with increasing  $n$ .

The important point is the fact that under pointwise convergence, continuity of functions of  $f_n(x)$  is not preserved. In fact,  $f_n(x)$  given in (3.4) is continuous for each  $n$  over the whole interval  $[0, 1]$ , whereas the limit  $f(x)$  given in (3.5) is discontinuous at  $x = 1$ . This indicates that interchanging the order of the limiting processes under pointwise convergence may produce different results, as expressed by

$$\lim_{x \rightarrow 1} \lim_{n \rightarrow \infty} f_n(x) \neq \lim_{n \rightarrow \infty} \lim_{x \rightarrow 1} f_n(x).$$

Similar phenomena might occur in connection with, integrability and differentiability of terms of functions  $f_n(x)$ . That is, under pointwise convergence, the limit of a sequence of integrable or differentiable functions may not be integrable or differentiable, respectively. Illustrative examples are given in Exercises 1 and 2 in Sect. 3.2.

### 3.2.2 Uniform Convergence

We know that if the sequence  $(f_n(x))$  is pointwise convergent to  $f(x)$  on  $x \in D$ , it is possible to choose  $N(x)$  for any small  $\varepsilon$  such that

$$m > N(x) \Rightarrow |f_m(x) - f(x)| < \varepsilon. \quad (3.6)$$

In general, the least value of  $N(x)$  that satisfies (3.6) will depend on  $x$ . But in certain cases, we can choose  $N$  *independent* of  $x$  such that  $|f_m(x) - f(x)| < \varepsilon$  for all  $m > N$  and *for all*  $x$  over the domain  $D$ . If this is true for any small  $\varepsilon$ , the sequence  $(f_n(x))$  is said to **converge uniformly** to  $f(x)$  on  $D$ . The formal definition is given below.

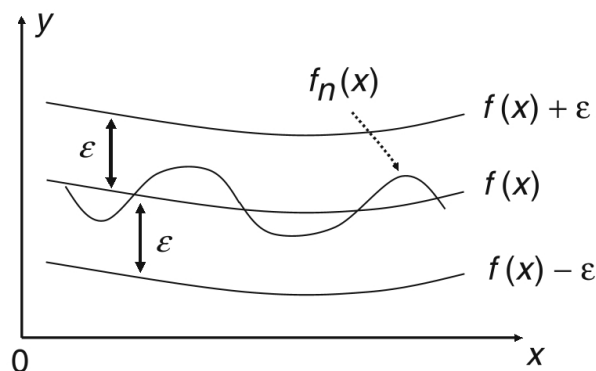
#### ♠ Uniform convergence:

The sequence  $(f_n)$  of real functions on  $D \subseteq \mathbf{R}$  **converges uniformly** to a function  $f$  on  $D$  if, given  $\varepsilon > 0$ , there is a positive integer  $N = N(\varepsilon)$  (which depends on  $\varepsilon$ ) such that

$$n > N \Rightarrow |f_n(x) - f(x)| < \varepsilon \text{ for all } x \in D.$$

Emphasis is placed on the fact that the integer  $N = N(\varepsilon, x)$  in the pointwise convergence *depends* on  $x$  in general, whereas  $N = N(\varepsilon)$  in the uniform convergence is *independent* of  $x$ . Under uniform convergence, therefore, by

taking  $n$  large enough we can always force the graph of  $y = f_n(x)$  into a band of width less than  $2\varepsilon$  centered around the graph of  $y = f(x)$  over the whole domain  $D$  (see Fig. 3.4).



**Fig. 3.4.** A function  $y = f_n(x)$  contained overall within a band of width less than  $2\varepsilon$

The definition of uniform convergence noted above is equivalent to the following statement.

♠ **Theorem:**

The sequence  $(f_n)$  of real functions on  $D \subseteq \mathbf{R}$  converges uniformly to  $f$  on  $D$  if and only if

$$\sup_{x \in D} |f_n(x) - f(x)| = 0 \text{ as } n \rightarrow \infty.$$

### 3.2.3 Cauchy Criterion for Series of Functions

As in the case of real sequences, the Cauchy criterion is available for testing uniform convergence for sequences of functions.

♠ **Cauchy criterion for uniform convergence:**

The sequence of  $f_n$  defined on  $D \subseteq \mathbf{R}$  converges uniformly to  $f$  on  $D$  if and only if, given  $\varepsilon > 0$ , there is a positive integer  $N = N(\varepsilon)$  such that

$$m, n > N \Rightarrow |f_m(x) - f_n(x)| < \varepsilon \text{ for all } x \in D, \quad (3.7)$$

or equivalently,

$$m, n > N \Rightarrow \sup_{x \in D} |f_m(x) - f_n(x)| < \varepsilon.$$

*Proof* Suppose that  $f_n(x)$  converges uniformly to  $f(x)$  on  $D$ . Let  $\varepsilon > 0$  and choose  $N \in \mathbf{N}$  such that

$$n > N \Rightarrow |f_n(x) - f(x)| < \frac{\varepsilon}{2} \text{ for all } x \in D.$$

If  $m, n \geq N$ , we have

$$|f_n(x) - f_m(x)| \leq |f_n(x) - f(x)| + |f(x) - f_m(x)| < \varepsilon \text{ for all } x \in D.$$

This result implies that if  $f_n(x)$  is uniformly convergent to  $f(x)$  on  $D$ , there exists an  $N$  that satisfies (3.7) for any small  $\varepsilon$ .

Next we consider the converse. Suppose that  $(f_n)$  satisfies the criterion given by (3.7). Then, for each point of  $x \in D$ ,  $(f_n(x))$  forms a Cauchy sequence and thus converges pointwise to

$$f(x) = \lim_{n \rightarrow \infty} f_n(x) \text{ for all } x \in D.$$

We now show that this convergence is uniform. Let  $n > N$  be fixed and take the limit  $m \rightarrow \infty$  in (3.7) to obtain

$$n > N \Rightarrow |f_n(x) - f(x)| < \varepsilon \text{ for all } x \in D,$$

where  $N$  is independent of  $x$ , from which we conclude that the convergence of  $(f_n)$  to  $f$  is uniform. ♣

### 3.2.4 Continuity of the Limit Function

The most important feature of uniform convergence is that it overcomes some of the shortcomings of pointwise convergence demonstrated in Sect. 3.2.1; i.e., pointwise convergence does not preserve continuity, integrability, and differentiability of terms of the functions  $f_n(x)$ . We now examine the situation under uniform convergence, starting with the continuity of  $f_n(x)$ .

#### ♠ Theorem:

If  $f_n$  converges uniformly to  $f$  on  $D \subseteq \mathbf{R}$ , then, if  $f_n$  is continuous at  $c \in D$ , so is  $f$ .

**Remark.** Note that the uniform convergence of  $f_n$  on  $D$  is a sufficient, but not a necessary, condition for  $f$  to be continuous. In fact, if  $f_n$  is not uniformly convergent on  $D$ , then its limit  $f$  may or may not be continuous at  $c \in D$ .

For the proof, it suffices to see that

$$\lim_{x \rightarrow c} f(x) = \lim_{x \rightarrow c} \lim_{n \rightarrow \infty} f_n(x) = \lim_{n \rightarrow \infty} \lim_{x \rightarrow c} f_n(x) = \lim_{n \rightarrow \infty} f_n(c) = f(c), \quad (3.8)$$

which guarantees the continuity of the limit function  $f(x)$  at  $x = c$ . In (3.8), we have used the interchangeability of limiting processes expressed by

$$\lim_{x \rightarrow c} \lim_{n \rightarrow \infty} f_n(x) = \lim_{n \rightarrow \infty} \lim_{x \rightarrow c} f_n(x),$$

which follows from the lemma below.

♠ **Lemma:**

Let  $c$  be a limit point of  $D \subseteq \mathbf{R}$  and assume that  $f_n$  converges uniformly to  $f$  on  $D \setminus \{c\}$ . If

$$\lim_{x \rightarrow c} f_n(x) = \ell_n \quad (3.9)$$

exists for each  $n$ , then

- (i)  $(\ell_n)$  is convergent, and
- (ii)  $\lim_{x \rightarrow c} f(x)$  exists and coincides with  $\lim_{n \rightarrow \infty} \ell_n$ ; i.e.,

$$\lim_{n \rightarrow \infty} \lim_{x \rightarrow c} f_n(x) = \lim_{x \rightarrow c} \lim_{n \rightarrow \infty} f_n(x). \quad (3.10)$$

*Proof* Let  $\varepsilon > 0$ . Since  $(f_n)$  converges uniformly on  $D \setminus \{c\}$ , it satisfies the Cauchy criterion; i.e., there is a positive integer  $N$  such that

$$m, n > N \Rightarrow |f_n(x) - f_m(x)| < \varepsilon \text{ for all } x \in D \setminus \{c\}. \quad (3.11)$$

Take the limit  $x \rightarrow c$  in (3.11) to obtain

$$m, n > N \Rightarrow |\ell_n - \ell_m| < \varepsilon. \quad (3.12)$$

This implies that  $(\ell_n)$  is a Cauchy sequence and thus convergent, which proves statement (i) above.

To prove (ii), let

$$\ell = \lim_{n \rightarrow \infty} \ell_n.$$

Set  $n = N$  and  $m \rightarrow \infty$  in (3.9), (3.11), and (3.12) to set the following results:

$$\lim_{x \rightarrow c} f_N(x) = \ell_N, \quad (3.13)$$

$$|f_N(x) - f(x)| < \varepsilon \text{ for all } x \in D \setminus \{c\}, \quad (3.14)$$

and

$$|\ell_N - \ell| < \varepsilon. \quad (3.15)$$

In addition, the existence of (3.13) implies that there exists a  $\delta > 0$  such that

$$|x - c| < \delta \text{ with } x \in D \setminus \{c\} \Rightarrow |f_N(x) - \ell_N| < \varepsilon. \quad (3.16)$$

Using (3.14), (3.15) and (3.16), we obtain

$$\begin{aligned} & |x - c| < \delta \text{ with } x \in D \setminus \{c\} \\ \Rightarrow & |f(x) - \ell| \leq |f(x) - f_N(x)| + |f_N(x) - \ell_N| + |\ell_N - \ell| < 3\varepsilon. \end{aligned}$$

This means that

$$\lim_{x \rightarrow c} f(x) = \ell,$$

which is equivalent to the desired result of (3.10). ♣

**Remark.** The contraposition of the theorem tells us that if the limit function  $f$  is discontinuous, the convergence of  $f_n$  is not uniform. The example in Sect. 3.2.1 demonstrated such a sequence.

### 3.2.5 Integrability of the Limit Function

We know that the limit function  $f(x)$  becomes continuous if the sequence  $(f_n(x))$  of continuous functions is uniformly convergent. This immediately results in the following theorem.

#### ♠ Theorem:

Suppose  $f_n$  be integrable on  $[a, b]$  for each  $n$ . Then, if  $f_n$  converges uniformly to  $f$  on  $[a, b]$ , the limit function  $f$  is also integrable, so that

$$\int_a^b f(x) dx = \lim_{n \rightarrow \infty} \int_a^b f_n(x) dx, \quad (3.17)$$

or equivalently,

$$\int_a^b \lim_{n \rightarrow \infty} f_n(x) dx = \lim_{n \rightarrow \infty} \int_a^b f_n(x) dx.$$

*Proof* Since  $f_n$  for every  $n$  is integrable on  $[a, b]$ , it is continuous (piecewise, at least) on  $[a, b]$ . Thus  $f(x)$  is also continuous (piecewise at least) on  $[a, b]$  in view of the theorem given in Sect. 3.2.4, so that  $f(x)$  is integrable on  $[a, b]$ . Furthermore, we observe that

$$\begin{aligned} \left| \int_a^b f_n(x) dx - \int_a^b f(x) dx \right| & \leq \int_a^b |f_n(x) - f(x)| dx \\ & \leq \int_a^b \sup_{x \in [a, b]} |f_n(x) - f(x)| dx \\ & \leq (b - a) \sup_{x \in [a, b]} |f_n(x) - f(x)|. \end{aligned}$$

The uniform convergence of  $(f_n)$  ensures that

$$\sup_{x \in [a, b]} |f_n(x) - f(x)| \rightarrow 0 \text{ as } n \rightarrow \infty,$$

which immediately gives the desired result shown in (3.17). ♣

*Remark.*

1. Note again that uniform convergence is a *sufficient* but not a *necessary* condition for (3.17) to be valid, so (3.17) may be valid even in the absence of uniform convergence. For instance, the convergence of  $(f_n)$  with  $f_n(x) = x^n$  on  $[0, 1]$  is not uniform but we have

$$\int_0^1 f_n(x) dx = \int_0^1 x^n dx = \frac{1}{n+1} \rightarrow 0 = \int_0^1 f(x) dx.$$

2. The conditions on  $f_n$  stated in the theorem will be significantly relaxed when we take up the Lebesgue integral in Chap. 6.

### 3.2.6 Differentiability of the Limit Function

After the last two subsections, readers may expect that results for differentiability will be similar to those for continuity and integrability; i.e., they may be tempted to conclude that the differentiability of terms of functions  $f_n(x)$  will be preserved if  $(f_n)$  converges uniformly to  $f$ . However, this is not the case. In fact, even if  $f_n$  converges uniformly to  $f$  on  $[a, b]$  and  $f_n$  is differentiable at  $c \in [a, b]$ , it may occur that

$$\lim_{n \rightarrow \infty} f'_n(c) \neq f'(c).$$

Consider the following example:

*Examples* Suppose the sequence  $(f_n)$  is defined by

$$f_n(x) = \sqrt{x^2 + \frac{1}{n^2}}, \quad x \in [-1, 1]. \quad (3.18)$$

Clearly (3.18) is differentiable for each  $n$ , and the sequence  $(f_n)$  converges uniformly on  $[-1, 1]$  to

$$f(x) = |x| \quad (3.19)$$

since

$$\begin{aligned} |f_n(x) - f(x)| &= \sqrt{x^2 + \frac{1}{n^2}} - \sqrt{x^2} \\ &= \frac{\frac{1}{n^2}}{\sqrt{x^2 + \frac{1}{n^2}} + \sqrt{x^2}} \leq \frac{1}{n} \rightarrow 0, \quad \text{for all } x \in [-1, 1]. \end{aligned}$$

However, the limit function  $f$  of (3.19) is not differentiable at  $x = 0$ . Hence, the desired result

$$\lim_{n \rightarrow \infty} f'_n(x) = f'(x) \quad (3.20)$$

breaks down at  $x = 0$ .

The following theorem provides sufficient conditions for (3.20) to be satisfied. The important point is that it requires the uniform convergence of the *derivatives*  $f'_n$ , not of the functions  $f_n$  themselves.

♠ **Theorem:**

Suppose  $(f_n)$  to be a sequence of differentiable functions on  $[a, b]$  that converge at a certain point  $x_0 \in [a, b]$ . If the sequence  $(f'_n)$  is uniformly convergent on  $[a, b]$ , then

- (i)  $(f_n)$  is also uniformly convergent on  $[a, b]$  to  $f$ ,
- (ii)  $f$  is differentiable on  $[a, b]$ , and
- (iii)  $\lim_{n \rightarrow \infty} f'_n(x) = f'(x)$ .

*Proof* Let  $\varepsilon > 0$ . From the convergence of  $(f_n(x_0))$  and the uniform convergence of  $(f'_n)$ , we conclude that there is an  $N \in \mathbf{N}$  such that

$$m, n > N \Rightarrow |f'_n(x) - f'_m(x)| < \varepsilon \text{ for all } x \in [a, b] \quad (3.21)$$

and

$$m, n > N \Rightarrow |f_n(x_0) - f_m(x_0)| < \varepsilon. \quad (3.22)$$

Given any two points  $x, t \in [a, b]$ , it follows from the mean value theorem applied to  $f_n - f_m$  that there is a point  $c$  between  $x$  and  $t$  such that

$$f_n(x) - f_m(x) - [f_n(t) - f_m(t)] = (x - t) [f'_n(c) - f'_m(c)].$$

Using (3.21), we have

$$m, n > N \Rightarrow |f_n(x) - f_m(x) - [f_n(t) - f_m(t)]| < \varepsilon |x - t|. \quad (3.23)$$

From (3.22) and (3.23), it follows that

$$\begin{aligned} |f_n(x) - f_m(x)| &\leq |f_n(x) - f_m(x) - [f_n(x_0) - f_m(x_0)]| + |f_n(x_0) - f_m(x_0)| \\ &< \varepsilon |x - x_0| + \varepsilon \\ &< \varepsilon(b - a + 1) = C\varepsilon, \text{ for all } x \in [a, b], \end{aligned}$$

Which means that  $(f_n)$  converges uniformly to some limit  $f$ . Hence, statement (i) has been proven.

Next we consider the proofs of **(ii)** and **(iii)**. For any fixed point  $x \in [a, b]$ , define

$$f_n(t) = \frac{f_n(t) - f_n(x)}{t - x}, \quad t \in [a, b] \setminus \{x\}$$

and

$$g(t) = \frac{f(t) - f(x)}{t - x}, \quad t \in [a, b] \setminus \{x\}.$$

Clearly,  $f_n \rightarrow g$  as  $n \rightarrow \infty$ ; furthermore, if  $m, n \geq N$ , the result of (3.23) tells us that

$$|f_n(t) - f_m(t)| < \varepsilon \quad \text{for all } t \in [a, b] \setminus \{x\}.$$

Thus in view of the Cauchy criterion, we see that  $f_n$  converges uniformly to  $g$  on  $[a, b] \setminus \{x\}$ . Now we observe that

$$\lim_{t \rightarrow x} f_n(t) = f'_n(x) \quad \text{for all } n \in \mathbf{N}. \quad (3.24)$$

Then, uniform convergence of  $f_n$  ensures taking the limit of  $n \rightarrow \infty$  in (3.24) followed by interchanging the order of the limit processes, which yields

$$\lim_{n \rightarrow \infty} \lim_{t \rightarrow x} f_n(t) = \lim_{t \rightarrow x} g(t) = \lim_{t \rightarrow x} \frac{f(t) - f(x)}{t - x} = f'(x) = \lim_{n \rightarrow \infty} f'_n(x).$$

This proves that  $f$  is differentiable at  $x$  and that

$$f'(x) = \lim_{n \rightarrow \infty} f'_n(x). \quad \clubsuit$$

*Remark.* That the uniform convergence of  $(f'_n)$  is just sufficient, not necessary, is seen by considering the sequence

$$f_n(x) = \frac{x^{n+1}}{n+1}, \quad x \in (0, 1).$$

This converges uniformly to 0, and its derivative  $f'_n(x) = x^n$  also converges to 0. The conclusions **(i)**–**(iii)** given in the theorem above are thus all satisfied. But the convergence of  $(f'_n)$  is not uniform.

## Exercises

1. For the function

$$f_n(x) = nx(1 - x^2)^n, \quad x \in [0, 1],$$

check that an interchange of the order of the limiting process  $n \rightarrow \infty$  and integration gives different results.

**Solution:** The given function is integrable for each  $n$  so that

$$\begin{aligned}\int_0^1 f_n(x)dx &= n \int_0^1 x(1-x^2)^n dx = \left[ \frac{-n}{2(n+1)} (1-x^2)^{n+1} \right]_0^1 \\ &= \frac{n}{2(n+1)} \rightarrow \frac{1}{2}.\end{aligned}$$

On the other hand, the limit given by

$$f(x) = \lim_{n \rightarrow \infty} f_n(x) = 0 \quad \text{for all } x \in [0, 1]$$

yields  $\int_0^1 f(x)dx = 0$ . We thus conclude that

$$\lim_{n \rightarrow \infty} \int_0^1 f_n(x)dx \neq \int_0^1 \lim_{n \rightarrow \infty} f_n(x)dx;$$

i.e., interchanging the order of integration and limiting processes is not in general allowed under pointwise convergence. ♣

2. For  $f_n(x)$  given by

$$f_n(x) = \begin{cases} -1 & x < -\frac{1}{n}, \\ \sin\left(\frac{n\pi x}{2}\right) & -\frac{1}{n} < x < \frac{1}{n}, \\ 1 & x > \frac{1}{n}, \end{cases}$$

check the continuity of its limit  $f(x) = \lim_{n \rightarrow \infty} f_n(x)$  at  $x = 0$ .

**Solution:**  $f_n(x)$  is differentiable for any  $x \in \mathbf{R}$  for all  $n$ , and thus is continuous at  $x = 0$  for all  $n$ . However, its limit,

$$f(x) = \begin{cases} -1 & x < 0, \\ 0 & x = 0, \\ 1 & x > 0 \end{cases}$$

is not continuous at  $x = 0$ . Hence, for the sequence of functions  $\{f_n(x)\}$ , the order of the limiting process  $n \rightarrow \infty$  and the differentiation with respect to  $x$  is not interchangeable. ♣

3. Show that the sequence of functions  $(f_n(x))$  defined by

$$f_n(x) = nxe^{-nx} \tag{3.25}$$

converges uniformly to  $f(x) = 0$  on  $x > 0$ .

**Solution:** In view of the previous theorem, we show that

$$\sup\{f_n(x) : x \geq a\} = 0 \text{ as } n \rightarrow \infty,$$

where  $a > 0$ . To prove this, we consider the derivative

$$f_n'(x) = ne^{-nx}(1 - nx). \quad (3.26)$$

It follows from (3.26) that  $x = 1/n$  is the only critical point of  $f_n$ . Now we choose a positive integer  $N$  such that  $a > 1/N$ . Then, the function  $f_n$  for each  $n \geq N$  has no critical point on  $x \geq a$ , and is monotonically decreasing. Therefore, the maximum of  $f_n(x)$  is attained at  $x = a$  for any  $n > N$ , with the result that

$$\sup_{x \in [a, \infty)} f_n(x) = f_n(a) = nae^{-na} \rightarrow 0 \quad (n \rightarrow \infty).$$

This holds for any  $a > 0$ ; hence, we conclude that  $f_n$  converges uniformly to 0 on  $(0, \infty)$ , i.e., on  $x > 0$ . ♣

*Remark.* Note that the range of uniform convergence of (3.25) is the open interval  $(0, \infty)$ , not the closed one  $[0, \infty)$ . Since in the latter case we have

$$\sup_{x \in [0, \infty)} f_n(x) = f_n\left(\frac{1}{n}\right) = \frac{1}{e} \not\rightarrow 0,$$

it is clear that  $(f_n)$  does not converge uniformly on  $[0, \infty)$ .

## 3.3 Series of Real Functions

### 3.3.1 Series of Functions

We close this chapter by considering convergence properties of series of real-valued functions. Assume a sequence  $(f_n)$  of functions defined on  $D \subseteq \mathbf{R}$ . By analogy with series of real numbers, we can define a series of functions by

$$S_n(x) = \sum_{k=1}^n f_k(x), \quad x \in D,$$

which gives a sequence  $(S_n) = (S_1, S_2, \dots)$ .

As  $n$  increases, the sequence  $(S_n)$  may or may not converge to a finite value, depending on the feature of functions  $f_k(x)$  as well as the point  $x$  in question. If the sequence converges for each point  $x \in D$  (i.e., converges

pointwise on  $D$ ), then the limit of  $S_n$  is called the **sum of the infinite series of functions**  $f_k(x)$  and is denoted by

$$S(x) = \lim_{n \rightarrow \infty} S_n(x) = \sum_{k=1}^{\infty} f_k(x), \quad x \in D.$$

It is obvious that the convergence of the series  $S_n(x)$  implies the pointwise convergence  $\lim_{n \rightarrow \infty} f_n(x) = 0$  on  $D$ . A series  $(S_n)$  that does not converge at a point  $x \in D$  is said to diverge at that point.

Applied to series of functions, the Cauchy criterion for uniform convergence takes the following form:

♠ **Cauchy criterion for series of functions:**

The series  $S_n$  is uniformly convergent on  $D$  if and only if for every small  $\varepsilon > 0$ , there is a positive integer  $N$  such that

$$\begin{aligned} n > m > N \\ \Rightarrow |S_n(x) - S_m(x)| = \left| \sum_{k=m+1}^n f_k(x) \right| < \varepsilon \quad \text{for all } x \in D. \end{aligned}$$

Set  $n = m + 1$  in the above criterion to obtain

$$n > N \Rightarrow |f_n(x)| < \varepsilon \quad \text{for all } x \in D.$$

This results implies that the uniform convergence of  $f_n(x) \rightarrow 0$  on  $D$  is a necessary condition for the convergence of  $S_n(x)$  to be uniform on  $D$ . We will use this theorem when proving a more practical test for uniform convergence known as the Weierstrass  $M$ -test, which is presented in Sect. 3.3.3.

### 3.3.2 Properties of Uniformly Convergent Series of Functions

When a given series of functions  $\sum f_k(x)$  is uniformly convergent, the properties of the sum  $S(x)$  in terms of continuity, integrability, and differentiability can be easily inferred from the properties of the separate terms  $f_k(x)$ . In fact, applying the theorems given in Sects. 3.2.4–3.2.6 to the sequence  $(S_n)$  and using the linearity regarding the limiting process, integration, and differentiation, we obtain the parallel theorems shown below.

♠ **Continuity of the sum:**

Suppose  $f_k(x)$  to be continuous for each  $k$ . If the sequence  $(S_n)$  of the series

$$S_n(x) = \sum_{k=1}^n f_k(x)$$

converges uniformly to  $S(x)$ , then  $S(x)$  is also continuous, so that

$$\lim_{t \rightarrow x} S(t) = \lim_{t \rightarrow x} \sum_{k=1}^{\infty} f_k(t) = \sum_{k=1}^{\infty} \lim_{t \rightarrow x} f_k(t).$$

♠ **Integrability of the sum:**

Suppose  $f_k$  to be integrable on  $[a, b]$  for all  $k$ . If  $(S_n)$  converges uniformly to  $S$  on  $[a, b]$ , we have

$$\int_a^b S(x) dx = \int_a^b \sum_{k=1}^{\infty} f_k(x) dx = \sum_{k=1}^{\infty} \int_a^b f_k(x) dx.$$

♠ **Differentiability of the sum:**

Let  $f_k$  be differentiable on  $[a, b]$  for each  $k$  and suppose that  $(S_n)$  converges to  $S$  at some point  $x_0 \in [a, b]$ . If the series  $\sum f'_k$  is uniformly convergent on  $[a, b]$ , then  $S_n(x)$  is also uniformly convergent on  $[a, b]$  and the sum  $S(x)$  is differentiable on  $[a, b]$ , so that

$$\frac{d}{dx} S(x) = \frac{d}{dx} \left[ \sum_{k=1}^{\infty} f_k(x) \right] = \sum_{k=1}^{\infty} \frac{df_k(x)}{dx} \quad \text{for all } x \in [a, b].$$

Observe that the second and third theorems provide a sufficient condition for performing term-by-term integration and differentiation, respectively, of an infinite series of functions. Without uniform convergence, such term-by-term calculations do not work.

### 3.3.3 Weierstrass $M$ -test

The following is a very useful and simple test for the uniform convergence of a series of functions.

♠ **Weierstrass  $M$  test:** If there is a sequence of positive constants  $M_k$  for any  $x$  on the interval  $[a, b]$  such that

$$|f_k(x)| \leq M_k \tag{3.27}$$

and if the series

$$\sum_{k=0}^{\infty} M_k \tag{3.28}$$

converges, then the series of functions  $\sum_{k=0}^{\infty} f_k(x)$  converges uniformly on  $x \in [a, b]$ .

*Proof* Since (3.28) converges, it follows from the Cauchy criterion that for any  $\varepsilon > 0$  there exists a number  $N$  such that

$$n > m > N \Rightarrow \left| \sum_{k=0}^n M_k - \sum_{k=0}^m M_k \right| = \sum_{k=m}^n M_k < \varepsilon. \quad (3.29)$$

Furthermore, in view of the inequality rule for absolute values of sums and the relation (3.27), it follows that

$$\left| \sum_{k=m}^n f_k(x) \right| \leq \sum_{k=m}^n |f_k(x)| \leq \sum_{k=m}^n M_k \quad (3.30)$$

for all  $x \in [a, b]$ . Note that the left-hand term in (3.30) can be rewritten as

$$\left| \sum_{k=m}^n f_k(x) \right| = \left| \sum_{k=0}^n f_k(x) - \sum_{k=0}^m f_k(x) \right|. \quad (3.31)$$

From (3.29), (3.30), and (3.31), it follows that

$$n \geq m > N \Rightarrow \left| \sum_{k=0}^n f_k(x) - \sum_{k=0}^m f_k(x) \right| < \varepsilon \text{ for all } x \in [a, b],$$

which clearly indicates the uniform convergence of  $\sum f_k(x)$  on  $[a, b]$ . ♣

## Exercises

1. Determine the convergence of the series  $\sum_{k=0}^{\infty} x^k$ .

**Solution:** It obviously converges to  $1/(1-x)$  on the interval  $[-a, a]$  if  $0 < a < 1$ . We show that this convergence is uniform on  $[-a, a]$  for any  $0 < a < 1$ . A partial sum yields  $S_n(x) = \sum_{k=0}^n x^k = (1-x^{n+1})/(1-x)$ , so that

$$|S(x) - S_n(x)| = \frac{|x|^{n+1}}{|1-x|} \leq \frac{a^{n+1}}{1-a} \text{ for } |x| \leq a.$$

Since  $0 < a < 1$ , the last term decreases monotonically with  $n$ ; hence, for a given  $\varepsilon > 0$ , we can find an  $N$  such that  $n > N \Rightarrow a^{n+1}/(1-a) < \varepsilon$ . Clearly the value of  $N$  does not depend on  $x$ . Therefore, we conclude that the infinite series  $\sum x^k$  is uniformly convergent on  $[-a, a]$  with  $0 < a < 1$ . ♣

2. Determine the convergence of the series  $\sum_{k=0}^{\infty} (1-x)x^k$ .

**Solution:** This converges to

$$S(x) = \begin{cases} 1, & \text{for } 0 < x < 1 \\ 0, & \text{at } x = 1 \end{cases}$$

but not uniformly. Actually, we have

$$|S(x) - S_n(x)| = \begin{cases} x^n & 0 < x < 1 \\ 0 & x = 1 \end{cases}$$

and if  $\varepsilon = 1/4$ , for instance, the inequality  $x^n < 1/4$  ( $0 < x < 1$ ) is false for every fixed  $n$  because  $x^n \rightarrow 1$  as  $x \rightarrow 1$ . ♣

3. Examine the uniform convergence of the series  $\sum_{k=1}^{\infty} f_k(x)$ , where  
 (i)  $f_k(x) = \frac{\cos kx}{k^2}$ , (ii)  $f_k(x) = \sin\left(\frac{x}{k^2}\right)$ , and (iii)  $f_k(x) = \frac{1}{k^2 x^2}$ .

**Solution:**

- (i) The series converges uniformly for every real  $x$ . Check this by taking  $M_k = 1/k^2$ .  
 (ii) Let  $D$  be a subset of  $\mathbf{R}$  bounded by  $c$ , i.e.,  $|x| \leq c$  for all  $x \in D$ . Then we have

$$\left| \sin\left(\frac{x}{k^2}\right) \right| \leq \frac{x}{k^2} \leq \frac{c}{k^2} \quad \text{for all } x \in D.$$

Taking  $M_k = c/k^2$  and noting that  $\sum M_k$  is convergent, we conclude that  $\sum f_k$  is uniformly convergent on any *bounded subset* of  $\mathbf{R}$ . Notably, however, this uniform convergence disappears when we extend the domain  $D$  to the whole  $\mathbf{R}$ . This is seen by noting that  $f_k \rightarrow 0$  pointwise on  $\mathbf{R}$ , but

$$\sup_{x \in \mathbf{R}} |f_k(x)| \geq \left| \sin\left(\frac{k^2 \pi/2}{k^2}\right) \right| = 1 \not\rightarrow 0,$$

which means that the convergence of  $(f_k)$  to 0 is not uniform on  $\mathbf{R}$ . In view of the theorem in Sect. 3.3.1, therefore, the series  $\sum f_k$  fails to converge uniformly on  $\mathbf{R}$ .

- (iii) The series  $\sum_k 1/(k^2 x^2)$  clearly converges pointwise on the open set  $\mathbf{R} \setminus \{0\}$ . Now let  $c > 0$ . For all  $x \in \mathbf{R}$  such that  $|x| > c$ , we have  $|f_k(x)| \leq 1/(k^2 c^2)$  for all  $k$ . Since  $\sum_k 1/(k^2 c^2)$  is

convergent, the series  $\sum f_k$  converges uniformly, by the  $M$ -test, on the closed set  $\mathbf{R} \setminus (-c, c) = (-\infty, -c] \cup [c, \infty)$  for all  $c > 0$ . But, although  $f_k \rightarrow 0$  pointwise on  $\mathbf{R} \setminus \{0\}$ , we have  $\sup_{x \neq 0} |f_k(x)| \geq |f_k(1/k)| = 1 \not\rightarrow 0$ . Hence,  $(f_k)$  does not converge uniformly to 0 on  $\mathbf{R} \setminus \{0\}$ , so the series  $\sum f_k$  does not converge uniformly on  $\mathbf{R} \setminus \{0\}$ . ♣

## 3.4 Improper Integrals

### 3.4.1 Definitions

Suppose that a given function  $f(x)$  is integrable on every open subinterval of  $(a, b)$ . We try to perform the integration  $\int_a^b f(x)dx$  under the following conditions:

1.  $f(x)$  is unbounded in a neighborhood of  $x = a$  or  $x = b$ .
2. The interval  $(a, b)$  itself is unbounded.

In Case 1, we define a definite integral,

$$\int_a^b f(x)dx = \lim_{X \rightarrow b-0} \int_a^X f(x)dx,$$

if  $f(x)$  is bounded and integrable on every finite interval  $(a, X)$  for  $a < X < b$ . Similarly, if  $f(x)$  is bounded and integrable on every  $(X, b)$  for  $a < X < b$ , we can define

$$\int_a^b f(x)dx = \lim_{X \rightarrow a+0} \int_X^b f(x)dx.$$

These definite integrals are called **improper integrals**. Straightforward extensions of these results to Case 2 yields the other improper integrals:

$$\int_a^\infty f(x)dx = \lim_{X \rightarrow \infty} \int_a^X f(x)dx$$

and

$$\int_{-\infty}^b f(x)dx = \lim_{X \rightarrow \infty} \int_{-X}^b f(x)dx.$$

*Examples* 1. The improper integral  $\int_1^\infty \frac{dx}{x^2}$  has the value 1 since

$$\int_1^\infty \frac{dx}{x^2} = \lim_{A \rightarrow \infty} \int_1^A \frac{dx}{x^2}.$$

2. The improper integral  $\int_0^4 \frac{dx}{\sqrt{x}}$  has the value 1 since

$$\int_0^4 \frac{dx}{\sqrt{x}} = \lim_{\varepsilon \rightarrow +0} \int_\varepsilon^4 \frac{dx}{\sqrt{x}} = \lim_{\varepsilon \rightarrow +0} \frac{2 - \sqrt{\varepsilon}}{2} = 1.$$

### 3.4.2 Convergence of an Improper Integral

An improper integral over  $f(x)$  is said to **converge** if and only if the corresponding limit exists. Furthermore, it is said to **converge absolutely** if and only if the corresponding improper integral over  $|f(x)|$  converges. (Keep in mind that absolute convergence implies convergence in the ordinary sense.) A convergent improper integral that does not converge absolutely is **conditionally convergent**.

An improper integral  $\int_a^b f(x, y)dx$  **converges uniformly** on a set  $S$  of values of  $y$  if and only if the corresponding limit converges uniformly on  $S$ . A relevant theorem is given below.

#### ♠ Continuity theorem

If  $f(x, y)$  is a continuous function, then  $\int_a^b f(x, y)dx$  is a continuous function of  $y$  in every open interval where the integral converges uniformly.

### 3.4.3 Principal Value Integral

Suppose that a bounded or unbounded open or closed interval,  $(a, b)$  or  $[a, b]$ , contains a discrete set of points  $x = c_1, c_2, \dots$ , such that  $f(x)$  is unbounded in a neighborhood of  $x = c_i$  ( $i = 1, 2, \dots$ ). Then, the integral  $\int_a^b f(x)dx$  may be defined as a sum of improper integrals, introduced in the previous subsection; i.e.,

$$\int_a^b f(x)dx = \lim_{X_1 \rightarrow a+0} \int_{X_1}^c f(x)dx + \lim_{X_2 \rightarrow b-0} \int_c^{X_2} f(x)dx \quad (a < c < b), \quad (3.32)$$

$$\int_a^b f(x)dx = \lim_{X_1 \rightarrow c-0} \int_a^{X_1} f(x)dx + \lim_{X_2 \rightarrow c+0} \int_{X_2}^b f(x)dx \quad (a < c < b), \quad (3.33)$$

$$\int_{-\infty}^{\infty} f(x)dx = \lim_{X_1 \rightarrow \infty} \int_{-X_1}^c f(x)dx + \lim_{X_2 \rightarrow \infty} \int_c^{X_2} f(x)dx \quad (3.34)$$

if the limits exists.

Even though the integrals (3.32), (3.33) and (3.34) do not exist, the limits of integrals

$$\lim_{x \rightarrow \infty} \int_{-X}^X f(x)dx \quad \text{and} \quad \lim_{\delta \rightarrow 0} \left[ \int_a^{c-\delta} f(x)dx + \int_{c+\delta}^b f(x)dx \right]$$

may exist. If any of these limits exist, the corresponding integral, (3.32), (3.33) or (3.34), is necessarily equal to its **principal value integral** (see Sect. 9.4.1).

### 3.4.4 Conditions for Convergence

In what follows we give the convergence criteria for improper integrals of the form

$$\int_a^\infty f(x)dx$$

and

$$\int_a^b f(x)dx = \lim_{X \rightarrow b-0} \int_a^X f(x)dx.$$

We assume that  $f(x)$  is bounded and integrable on every bounded interval  $(a, X)$  that does not contain the upper limit of integration.

♠ **Cauchy's test** (= necessary and sufficient conditions for convergence):

The improper integral  $\int_a^\infty f(x)dx$  converges if and only if for every positive real number  $\varepsilon$ , there exists a real number  $M > a$  such that

$$X_2 > X_1 > M \Rightarrow \left| \int_{X_1}^{X_2} f(x)dx \right| < \varepsilon.$$

Similarly,  $\int_a^\infty f(x)dx$  converges if and only if for every positive  $\varepsilon$ , there exists a positive  $\delta < b - a$  such that

$$b - X_2 < b - X_1 < \delta \Rightarrow \left| \int_{X_1}^{X_2} f(x)dx \right| < \varepsilon.$$

Necessary and sufficient conditions for an improper integral to converge uniformly are stated below.

♠ **Weierstrass test**

The improper integral  $\int_a^\infty f(x, y)dx$  [or  $\int_a^b f(x, y)dx$ ] converges uniformly and absolutely on every set  $S$  of values of  $y$  such that  $|f(x, y)| \leq g(x)$  on the interval of integration, where  $g(x)$  is a real comparison function whose integral  $\int_a^\infty g(x)dx$  [or  $\int_a^b g(x)dx$ , respectively] converges.

## Exercises

1. Show that the integral  $\int_{\pi}^{\infty} \frac{\sin x}{x} dx$  converges.

**Solution:** We have

$$\int_{\pi}^A \frac{\sin x}{x} dx \leq \left[ \frac{-\cos x}{x} \right]_{\pi}^A - \int_{\pi}^A \frac{\cos x}{x^2} dx,$$

so that

$$\left| \int_{\pi}^A \frac{\sin x}{x} dx \right| \leq \frac{1}{\pi} + \frac{1}{A} + \int_{\pi}^A \frac{dx}{x^2} = \left( \frac{1}{\pi} + \frac{1}{A} \right) + \left( \frac{1}{\pi} - \frac{1}{A} \right) = 2\pi.$$

This completes the proof. ♣

2. Show that  $\int_1^{\infty} \left| \frac{\sin x}{x} \right| dx$  diverges.

**Solution:** It follows that

$$\begin{aligned} \int_{n\pi}^{(n+1)\pi} \left| \frac{\sin x}{x} \right| dx &= \int_0^{\pi} \frac{\sin x}{n\pi + x} dx > \frac{1}{(n+1)\pi} \int_0^{\pi} \sin x dx \\ &= \frac{2}{(n+1)\pi} > \frac{2}{\pi} \int_{n+1}^{n+2} \frac{dx}{x}. \end{aligned}$$

Hence, for  $n > 1$  we have

$$\int_{\pi}^{n\pi} \left| \frac{\sin x}{x} \right| dx > \frac{2}{\pi} \int_2^{n+1} \frac{dx}{x} = \frac{1}{\pi} \log(n+1) \rightarrow \infty, \quad (n \rightarrow \infty). \quad \clubsuit$$

3. Suppose that  $f(x)$  is continuous within an interval  $(a, b]$  and diverges at  $x = a$ . Prove that  $\int_a^b f(x) dx$  converges if  $(x-a)^p |f(x)|$  is bounded on the interval for  $0 < p < 1$ .

**Solution:** We assume that there is an appropriate positive number  $M$  such that

$$(x-a)^p |f(x)| < M \quad \text{for all } x \in (a, b].$$

Then we obtain

$$\begin{aligned} \int_{a+\varepsilon}^b |f(x)| dx &< M \int_{a+\varepsilon}^b \frac{dx}{(x-a)^p} = M \left[ \frac{(x-a)^{1-p}}{1-p} \right]_{a+\varepsilon}^b \\ &= \frac{M}{1-p} [(b-a)^{1-p} - \varepsilon^{1-p}] < \frac{M}{1-p} (b-a)^{1-p} \\ &\quad (\text{since } 1-p > 0). \end{aligned} \tag{3.35}$$

Note that the integral on the left-hand side of (3.35) is monotonically increasing with decreasing  $\varepsilon$ , since  $|f(x)| \geq 0$  over the integration interval. Yet it is bounded from above, as proved in (3.35). Hence, we conclude that the given integral is convergent (absolutely). ♣

4. Suppose that  $f(x)$  is continuous within  $[a, \infty)$  and that  $x^p|f(x)|$  is bounded there for  $p > 1$ . Show that the integral  $\int_a^b f(x)dx$  converges.

**Solution:** It follows from hypothesis that there is a positive number  $M$  such that

$$x^p|f(x)| < M \quad \text{for all } x \geq a.$$

Hence, we have for any  $X > a$ ,

$$\int_a^X |f(x)|dx < M \int_a^X \frac{dx}{x^p} = \frac{-M}{p-1} \left[ \frac{1}{x^{p-1}} \right]_a^X < \frac{M}{p-1} \frac{1}{a^{p-1}},$$

which completes the proof. ♣

## Functional Analysis



## Hilbert Spaces

**Abstract** A Hilbert space is an abstract vector space with the following two properties: the inner product property (Sect. 4.1.3), which determines the geometry of the vector space, and the completeness property (Sect. 4.1.6), which guarantees the self-consistency of the space. Most of the mathematical topics covered in this volume are based on Hilbert spaces. In particular,  $L^p$  spaces and  $l^p$  spaces (Sect. 4.3), which are specific classes of Hilbert spaces, are crucial for the formulation of the theories of orthonormal polynomials, Lebesgue integrals, Fourier analyses, and others, as we discuss in subsequent chapters.

### 4.1 Hilbert Spaces

#### 4.1.1 Introduction

This section provides a framework for an understanding of **Hilbert spaces**. Plainly speaking, Hilbert spaces are the generalization of familiar finite-dimensional spaces to the infinite-dimensional case. In fact, the geometric structure of Hilbert spaces is very similar to that of ordinary Euclidean geometry. This analogy comes from the fact that the concept of **orthogonality** can be introduced in any Hilbert space so that the familiar Pythagorean theorem holds for elements involved in the space. Moreover, owing to its generality, a large number of problems in physics and engineering can be successfully treated with a geometric point of view in Hilbert spaces.

As we shall see later, Hilbert spaces are defined as a specific class of **vector spaces** endowed with the following two properties: **inner product** and **completeness**. The former property leads to a rich geometric structure and the latter enables us to describe an element in the space in terms of a set of **orthonormal bases**. These facts result in the possibility of establishing a wide variety of **complete orthonormal sets of functions** in Hilbert spaces; we discussed this point in detail in Sects. 5.1 and 5.2. For a better understanding of subsequent discussions, we provide all necessary definitions in

this section, and then describe several important consequences relevant to an understanding of the nature of Hilbert spaces.

### 4.1.2 Abstract Vector Spaces

In order to make this text self-contained, we first give a brief summary of the definition of **vector spaces**. A more precise description of vector spaces and some related matters will be provided in Sect. 4.2.1.

#### ♠ Vector spaces:

A vector space  $V$  is a collection of elements called **vectors**, which we denote by  $\mathbf{x}, \mathbf{y}, \dots$ , that satisfy the following postulates:

1. There exists an operation  $(+)$  on the vectors  $\mathbf{x}$  and  $\mathbf{y}$  such that  $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$ , where the resultant quantity  $\mathbf{y} + \mathbf{x}$  also must be a vector.
2. There exists an **identity vector** (denoted by  $\mathbf{0}$ ) that yields  $\mathbf{x} + \mathbf{0} = \mathbf{x}$ .
3. For every  $\mathbf{x} \in V$ , there exists a vector  $\alpha\mathbf{x} \in V$  in which  $\alpha$  is an arbitrary scalar (real and complex). In addition,

$$\begin{aligned}\alpha(\beta\mathbf{x}) &= (\alpha\beta)\mathbf{x}, & 1(\mathbf{x}) &= \mathbf{x} \text{ for all } \mathbf{x}, \\ \alpha(\mathbf{x} + \mathbf{y}) &= \alpha\mathbf{x} + \alpha\mathbf{y}, & (\alpha + \beta)\mathbf{x} &= \alpha\mathbf{x} + \beta\mathbf{x}.\end{aligned}$$

Emphasis is placed on the fact that vector spaces are not limited to a set of geometric *arrows* embedded in a **Euclidean space** (see Sects. 4.1.3 and 19.2.3); rather, they are general mathematical systems that have a specific algebraic structure. Several examples of such abstract vector spaces are given below.

*Examples* 1. The set of all  $n$ -tuples of complex numbers denoted by

$$\mathbf{x} = (\xi_1, \xi_2, \dots, \xi_n)$$

forms a vector space if the addition of vectors and the multiplication of a vector by a scalar are defined by

$$\begin{aligned}\mathbf{x} + \mathbf{y} &= (\xi_1, \xi_2, \dots, \xi_n) + (\eta_1, \eta_2, \dots, \eta_n) \\ &= (\xi_1 + \eta_1, \xi_2 + \eta_2, \dots, \xi_n + \eta_n), \\ \alpha\mathbf{x} &= \alpha(\xi_1, \xi_2, \dots, \xi_n) = (\alpha\xi_1, \alpha\xi_2, \dots, \alpha\xi_n).\end{aligned}$$

2. The set of all complex numbers  $\{z\}$  forms a **complex vector space** (see Sect. 4.2.1), where  $z_1 + z_2$  and  $\alpha z$  are interpreted as ordinary complex numerical addition and multiplication,
3. The set of all polynomials in a real variable  $x$ , constituting the set  $\{1, x, x^2, x^3, \dots\}$ , with complex coefficients is a complex vector space if vector addition and scalar multiplication are the ordinary addition of two polynomials and the multiplication of a polynomial by a complex number, respectively.

### 4.1.3 Inner Product

The structure of a vector space is enormously enriched by introducing the concept of **inner product**, which enables us to define the length of a vector in a given vector space or the angle between the two vectors involved.

♠ **Inner product:**

An inner product is a scalar-valued function of the ordered pair of vectors  $\mathbf{x}$  and  $\mathbf{y}$  such that

1.  $(\mathbf{x}, \mathbf{y}) = (\mathbf{y}, \mathbf{x})^*$ .
2.  $(\alpha\mathbf{x} + \beta\mathbf{y}, \mathbf{z}) = \alpha^*(\mathbf{x}, \mathbf{z}) + \beta^*(\mathbf{y}, \mathbf{z})$ , where  $\alpha$  and  $\beta$  are certain complex numbers.
3.  $(\mathbf{x}, \mathbf{x}) \geq 0$  for any  $\mathbf{x}$ ;  $(\mathbf{x}, \mathbf{x}) = 0$  if and only if  $\mathbf{x} = 0$ .

Here, the asterisk (\*) indicates that one is to take the **complex conjugate**.

**Remark.** Vector spaces endowed with an inner product are called **inner product spaces**. In particular, a real inner product space is called a **Euclidean space** and a complex inner product space is called a **unitary space**.

The algebraic properties **1** and **2** are in principle the same as those governing the **scalar product** in ordinary vector algebra in a real vector space. The only property that is not obvious is that in a complex space, the inner product is not linear, but rather **conjugate linear** with respect to the first factor, i.e.,

$$(\alpha\mathbf{x}, \mathbf{y}) = \alpha^*(\mathbf{x}, \mathbf{y}).$$

**Examples 1.** The simplest, but an important, example of an inner product space is the space, denoted by  $\mathbf{C}$ , that consists of a set of complex numbers  $\{z_1, z_2, \dots, z_n\}$ . For two vectors  $\mathbf{x} = (\xi_1, \xi_2, \dots, \xi_n)$  and  $\mathbf{y} = (\eta_1, \eta_2, \dots, \eta_n)$  on  $\mathbf{C}$ , the inner product is defined by

$$(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^n \xi_i^* \eta_i.$$

2. Suppose that  $f(x)$  and  $g(x)$  are polynomials in the complex vector space defined on the closed interval  $x \in [0, 1]$ . They then constitute an inner product space under the inner product defined by

$$(f, g) = \int_0^1 f(x)^* g(x) w(x) dx,$$

where  $w(x)$  is a **weight function**. The weight function becomes important when defining the inner product of polynomials, which is treated in Chap. 5.

- 3.** If  $\mathbf{x} = [\xi_1, \xi_2, \xi_3, \xi_4]$  and  $\mathbf{y} = [\eta_1, \eta_2, \eta_3, \eta_4]$  are column **four-vectors** having real-valued elements, then the quantity

$$(\mathbf{x}, \mathbf{y}) \equiv \xi_1\eta_1 + \xi_2\eta_2 + \xi_3\eta_3 - \xi_4\eta_4 \quad (4.1)$$

satisfies requirements **1** and **2** for an inner product, but not **3** since the quantity  $(\mathbf{x}, \mathbf{x})$  is *not* positive-definite. Thus the entity (4.1) is not an inner product, but it plays an important role in the theory of special relativity.

For a complex vector space, the inner product is not symmetrical as it is in a real vector space. That is,  $(\mathbf{x}, \mathbf{y}) \neq (\mathbf{y}, \mathbf{x})$  but rather  $(\mathbf{x}, \mathbf{y}) = (\mathbf{y}, \mathbf{x})^*$ . This implies that  $(\mathbf{x}, \mathbf{x})$  is real for every  $\mathbf{x}$ , so we can define the *length* of the vector  $\mathbf{x}$  by

$$\|\mathbf{x}\| = (\mathbf{x}, \mathbf{x})^{1/2}.$$

Since  $(\mathbf{x}, \mathbf{x}) \geq 0$ ,  $\|\mathbf{x}\|$  is always nonnegative and real. The quantity  $\|\mathbf{x}\|$  is referred to as the **norm** of the vector  $\mathbf{x}$ . Note also that

$$\|\alpha\mathbf{x}\| = (\alpha\mathbf{x}, \alpha\mathbf{x})^{1/2} = [\alpha^*\alpha(\mathbf{x}, \mathbf{x})]^{1/2} = |\alpha| \cdot \|\mathbf{x}\|.$$

**Remark.** Precisely speaking, the quantity  $\|\mathbf{x}\|$  introduced above is a special kind of norm that is associated with an inner product; in fact, the norm was originally a more general concept that was independent of the inner product (see Sect. 4.2.2).

#### 4.1.4 Geometry of Inner Product Spaces

Once a vector space is endowed with an inner product, several important theorems that can be easily interpreted in analogy with Euclidean geometry can be applied. The following three theorems characterize the geometric nature of inner product spaces ( $\mathbf{x} \neq 0$  and  $\mathbf{y} \neq 0$  are assumed; otherwise the theorems all become trivial).

##### ♠ Schwarz inequality:

For any two elements  $\mathbf{x}$  and  $\mathbf{y}$  of an inner product space, we have

$$|(\mathbf{x}, \mathbf{y})| \leq \|\mathbf{x}\| \|\mathbf{y}\|. \quad (4.2)$$

The equality holds if and only if  $\mathbf{x}$  and  $\mathbf{y}$  are linearly independent.

*Proof* From the definition of the inner product, we have

$$0 \leq (\mathbf{x} + \alpha\mathbf{y}, \mathbf{x} + \alpha\mathbf{y}) = (\mathbf{x}, \mathbf{x}) + \alpha(\mathbf{x}, \mathbf{y}) + \alpha^*(\mathbf{y}, \mathbf{x}) + |\alpha|^2(\mathbf{y}, \mathbf{y}). \quad (4.3)$$

Now, set  $\alpha = -(\mathbf{x}, \mathbf{y})/(\mathbf{y}, \mathbf{y})$  and multiply by  $(\mathbf{y}, \mathbf{y})$  to obtain

$$0 \leq (\mathbf{x}, \mathbf{x})(\mathbf{y}, \mathbf{y}) - |(\mathbf{x}, \mathbf{y})|^2,$$

which gives Schwarz's inequality.

Next, we prove the statement of the equality in (4.2). If  $\mathbf{x}$  and  $\mathbf{y}$  are linearly dependent, then  $\mathbf{y} = \alpha\mathbf{x}$  for some complex number  $\alpha$  so that we have

$$|(\mathbf{x}, \mathbf{y})| = |(\mathbf{x}, \alpha\mathbf{x})| = |\alpha|(\mathbf{x}, \mathbf{x}) = |\alpha|\|\mathbf{x}\| \|\mathbf{x}\| = \|\mathbf{x}\| \|\alpha\mathbf{x}\| = \|\mathbf{x}\| \|\mathbf{y}\|.$$

The converse is also true; let  $\mathbf{x}$  and  $\mathbf{y}$  be vectors such that  $|(\mathbf{x}, \mathbf{y})| = \|\mathbf{x}\| \|\mathbf{y}\|$ , or equivalently,

$$|(\mathbf{x}, \mathbf{y})|^2 = (\mathbf{x}, \mathbf{y})(\mathbf{y}, \mathbf{x}) = (\mathbf{x}, \mathbf{x})(\mathbf{y}, \mathbf{y}) = \|\mathbf{x}\|^2 \|\mathbf{y}\|^2. \quad (4.4)$$

Then we set

$$\begin{aligned} & \|(\mathbf{y}, \mathbf{y})\mathbf{x} - (\mathbf{y}, \mathbf{x})\mathbf{y}\|^2 \\ &= \|\mathbf{y}\|^4 \|\mathbf{x}\|^2 + |(\mathbf{y}, \mathbf{x})|^2 \|\mathbf{y}\|^2 - \|\mathbf{y}\|^2 (\mathbf{y}, \mathbf{x})(\mathbf{x}, \mathbf{y}) - \|\mathbf{y}\|^2 (\mathbf{y}, \mathbf{x})^*(\mathbf{y}, \mathbf{x}) \\ &= 0, \end{aligned} \quad (4.5)$$

where the postulate (4.4) and the relation  $(\mathbf{y}, \mathbf{x})^* = (\mathbf{x}, \mathbf{y})$  were used. The result (4.5) means that

$$(\mathbf{y}, \mathbf{y})\mathbf{x} - (\mathbf{x}, \mathbf{y})\mathbf{y} = \mathbf{0},$$

which clearly shows that  $\mathbf{x}$  and  $\mathbf{y}$  are linearly dependent, which completes the proof. ♣

#### ♠ Triangle inequality:

For any two elements  $\mathbf{x}$  and  $\mathbf{y}$  of an inner product space, we have

$$\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|.$$

*Proof* Setting  $\alpha = 1$  in (4.3), we have

$$\begin{aligned} \|\mathbf{x} + \mathbf{y}\|^2 &= (\mathbf{x}, \mathbf{x}) + (\mathbf{y}, \mathbf{y}) + 2\operatorname{Re}(\mathbf{x}, \mathbf{y}) \\ &\leq (\mathbf{x}, \mathbf{x}) + (\mathbf{y}, \mathbf{y}) + 2|(\mathbf{x}, \mathbf{y})| \\ &\leq \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 + 2\|\mathbf{x}\|\|\mathbf{y}\| \quad (\text{by Schwarz's inequality}) \\ &= (\|\mathbf{x}\| + \|\mathbf{y}\|)^2, \end{aligned}$$

which proves the desired inequality. ♣

♠ **Parallelogram law:**

For any two elements  $\mathbf{x}$  and  $\mathbf{y}$  of an inner product space, we have

$$\|\mathbf{x} + \mathbf{y}\|^2 + \|\mathbf{x} - \mathbf{y}\|^2 = 2(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2).$$

*Proof* We have

$$\begin{aligned}\|\mathbf{x} + \mathbf{y}\|^2 &= (\mathbf{x}, \mathbf{x}) + (\mathbf{x}, \mathbf{y}) + (\mathbf{y}, \mathbf{x}) + (\mathbf{y}, \mathbf{y}) \\ &= \|\mathbf{x}\|^2 + (\mathbf{x}, \mathbf{y}) + (\mathbf{y}, \mathbf{x}) + \|\mathbf{y}\|^2.\end{aligned}\tag{4.6}$$

Now replace  $\mathbf{y}$  by  $-\mathbf{y}$  to obtain

$$\|\mathbf{x} - \mathbf{y}\|^2 = \|\mathbf{x}\|^2 - (\mathbf{x}, \mathbf{y}) - (\mathbf{y}, \mathbf{x}) + \|\mathbf{y}\|^2.\tag{4.7}$$

By adding (4.6) and (4.7), we attain our objective. ♣

#### 4.1.5 Orthogonality

One of the most important consequences of having the inner product is being able to define the **orthogonality** of vectors. The orthogonality allows us to establish a set of **orthonormal bases** that span the inner product space in question, thus yielding a useful way to analyze both the nature of the space itself and the relation between the constituents involved in that space.

♠ **Orthogonality:**

Two vectors  $\mathbf{x}$  and  $\mathbf{y}$  in an inner product space are called **orthogonal** if and only if  $(\mathbf{x}, \mathbf{y}) = 0$ .

Notably, if  $(\mathbf{x}, \mathbf{y}) = 0$ , then  $(\mathbf{x}, \mathbf{y}) = (\mathbf{y}, \mathbf{x})^* = 0$  so that  $(\mathbf{y}, \mathbf{x}) = 0$  as well. Thus, the orthogonality is a symmetric relation, although the inner product is not symmetric. Note also that the **zero vector**  $\mathbf{0}$  is orthogonal to every vector in the inner product space.

A set of  $n$  vectors  $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$  is called **orthonormal** if  $(\mathbf{x}_i, \mathbf{x}_j) = \delta_{ij}$  for all  $i$  and  $j$ , where  $\delta_{ij}$  is the **Kronecker delta**. That is, the orthonormality of a set of vectors means that each vector is orthogonal to all the others in the set and is normalized to unit length.

It follows that any vector  $\mathbf{x}$  may be normalized by dividing by its length to form the new vector  $\mathbf{x}/\|\mathbf{x}\|$  with unit length. An example of an orthonormal set of vectors is the set of three unit vectors,  $\{\mathbf{e}_i\}$  ( $i = 1, 2, 3$ ), for the three-dimensional **Cartesian space**.

The following theorem is important in various fields of mathematical physics.

♠ **Theorem:**

An orthonormal set is linearly independent.

(Proof of the theorem is given in Exercise 1). Importantly, the above theorem suggests that any orthonormal set serves as a **basis** for an inner product space of interest (see Sect. 4.2.5). Below is another consequence of the orthonormal set of vectors; its proof is given in Exercise 2.

♠ **Bessel inequality:**

If  $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$  is a set of orthonormal vectors and  $\mathbf{x}$  is any vector defined in the same inner product space, then

$$\|\mathbf{x}\|^2 \geq \sum_i |r_i|^2, \quad (4.8)$$

where  $r_i = (\mathbf{x}_i, \mathbf{x})$ . Furthermore, the vector  $\mathbf{x}' = \mathbf{x} - \sum_i r_i \mathbf{x}_i$  is orthogonal to each  $\mathbf{x}_j$ .

#### 4.1.6 Completeness of Vector Spaces

Having described features of inner product spaces, we turn now to another important concept relevant to the nature of Hilbert spaces, i.e., **completeness**. When a vector space is *finite* dimensional, the completeness of an orthonormal set involved in the space may be characterized by the fact that it is not contained in any larger orthonormal set. (This is intuitively understood by considering the **Cartesian basis**  $\mathbf{e}_i$  ( $i = 1, 2, 3$ ) in a three-dimensional Euclidean space.) When considering an *infinite*-dimensional space, however, the completeness must be determined via the Cauchy criterion, which we discussed in Sect. 2.2. The following is a preliminary definition

♠ **Cauchy sequence of vectors:**

A sequence  $\{\mathbf{x}_1, \mathbf{x}_2, \dots\}$  of vectors is called a **Cauchy sequence** of vectors if for any positive  $\varepsilon > 0$ , there exists an appropriate number  $N$  such that  $\|\mathbf{x}_m - \mathbf{x}_n\| < \varepsilon$  for all  $m, n > N$ .

In plain words, a sequence is a Cauchy sequence if the terms  $\mathbf{x}_m$  and  $\mathbf{x}_n$  in the sequence come closer and closer to each other as  $m, n \rightarrow \infty$

♠ **Convergence of a sequence of vectors:**

A sequence  $\{\mathbf{x}_1, \mathbf{x}_2, \dots\}$  is said to be convergent if there exists an element  $\mathbf{x}$  such that  $\|\mathbf{x}_n - \mathbf{x}\| \rightarrow 0$ .

♠ **Completeness of a vector space:**

If every Cauchy sequence in a space is convergent, we say that the space is **complete**.

*Remark.* Here the norm  $\|\mathbf{x}\| = (\mathbf{x}, \mathbf{x})^{1/2}$  associated with an inner product is employed to define a Cauchy sequence, since we are focusing on inner product spaces. However, the concepts of Cauchy sequence and completeness both apply to more general vector spaces in which even a norm is unnecessary (see Sect. 4.1.6 for details).

#### 4.1.7 Several Examples of Hilbert Spaces

Now we are ready to define Hilbert spaces.

♠ **Hilbert space:**

If an inner product space is complete, it is called a **Hilbert space**.

*Examples 1.* Column-vector spaces with  $n$  real and complex components, denoted by  $\mathbf{R}^n$  and  $\mathbf{C}^n$ , respectively, are finite-dimensional Hilbert spaces if endowed with an inner product  $(\mathbf{x}, \mathbf{y}) \equiv \sum_{i=1}^n x_i^* y_i$ . Completeness can be proved using the Bolzano–Weierstrass theorem (see Appendix A).

**2.** Assume an infinite-dimensional vector  $\mathbf{x} = (x_1, x_2, \dots)$ , where  $x_i$  is a real or complex number satisfying the condition

$$\sum_{i=1}^{\infty} |x_i|^2 < \infty.$$

Then, vector spaces spanned by a set of vectors  $\{\mathbf{x}\}$ , called  $\ell^2$  **spaces** (see Sect. 4.3), are Hilbert spaces under the inner product

$$(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{\infty} x_i^* y_i.$$

Completeness will be proved in Sect. 4.3.1.

4. Assume a set of square-integrable functions  $f(x)$  expressed by

$$\int_a^b |f(x)|^2 dx < \infty.$$

Then, the collection of all square-integrable functions, called the  $L^2$  **space**, is a Hilbert space endowed with the inner product

$$(f, g) = \int_a^b f(x)^* g(x) dx. \quad (4.9)$$

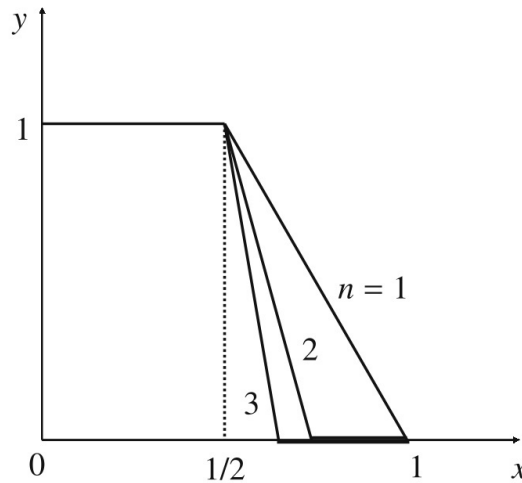
Completeness will be proved given in Sect. 4.3.2.

5. Finally we show an example of an *incomplete* inner product space. Assume the following sequence of real-valued continuous functions  $\{f_1(x), f_2(x), \dots\}$ , each of which is defined within the interval  $[0, 1]$ :

$$f_n(x) = \begin{cases} 1, & \text{for } 0 \leq x \leq \frac{1}{2}, \\ 1 - 2n(x - \frac{1}{2}) & \text{for } \frac{1}{2} \leq x \leq \frac{1}{2n} + \frac{1}{2}, \\ 0, & \text{for } \frac{1}{2n} + \frac{1}{2} \leq x \leq 1. \end{cases} \quad (4.10)$$

The graphs of  $f_n(x)$  for  $n = 1, 2, 3$  are given in Fig. 4.1. After some algebra, we obtain

$$\begin{aligned} \|f_n(x) - f_m(x)\| &= \left[ \int_0^1 (f_n - f_m)^2 dx \right]^{1/2} \\ &= \left(1 - \frac{n}{m}\right) \sqrt{\frac{1}{6n}} \rightarrow 0 \quad \text{as } m, n \rightarrow \infty \quad (m > n). \end{aligned}$$



**Fig. 4.1.** The function  $f_n(x)$  given in (4.10). The sequence  $\{f_n(x)\}$  converges to a step function in the limit of  $n \rightarrow \infty$

Thus,  $\{f_n\}$  is a Cauchy sequence owing to the inner product given by (4.9). However, this sequence converges to the limit function

$$f(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq \frac{1}{2}, \\ 0 & \text{if } \frac{1}{2} \leq x \leq 1, \end{cases}$$

which is not continuous and, hence, is not an element of the original inner product space. Consequently, the sequence is not complete, and thus is not a Hilbert space.

## Exercises

1. Show that an orthonormal set is linearly independent.

**Solution:** Recall that a set of vectors  $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$  is said to be **linearly independent** if and only if

$$\sum_{i=1}^n \alpha_i \mathbf{x}_i = 0 \Rightarrow \alpha_i = 0 \text{ for all } i.$$

Now suppose that a set  $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$  is orthonormal and satisfies the relation  $\sum_i \alpha_i \mathbf{x}_i = 0$ . Then, for any  $j$ , the orthonormal condition  $(\mathbf{x}_i, \mathbf{x}_j) = \delta_{ij}$  results in

$$0 = \left( \mathbf{x}_j, \sum_{i=1}^n \alpha_i \mathbf{x}_i \right) = \sum_{i=1}^n \alpha_i (\mathbf{x}_j, \mathbf{x}_i) = \sum_i \alpha_i \delta_{ij} = \alpha_j.$$

Therefore, the set is linearly independent. ♣

2. Show the Bessel inequality for  $\mathbf{x}$  given by 4.8 and the orthogonality of the vector  $\mathbf{x}' = \mathbf{x} - \sum_i r_i \mathbf{x}_i$  to each  $\mathbf{x}_j$ .

**Solution:** We consider the inequality

$$\begin{aligned} 0 \leq \|\mathbf{x}'\|^2 &= (\mathbf{x}', \mathbf{x}') = \left( \mathbf{x} - \sum_{i=1}^n r_i \mathbf{x}_i, \mathbf{x} - \sum_{j=1}^n r_j \mathbf{x}_j \right) \\ &= (\mathbf{x}, \mathbf{x}) - \sum_{i=1}^n r_i^* (\mathbf{x}_i, \mathbf{x}) - \sum_{j=1}^n r_j (\mathbf{x}, \mathbf{x}_j) + \sum_{i,j=1}^n r_i^* r_j (\mathbf{x}_i, \mathbf{x}_j) \\ &= \|\mathbf{x}\|^2 - \sum_{i=1}^n |r_i|^2 - \sum_{j=1}^n |r_j|^2 + \sum_{j=1}^n |r_j|^2 \\ &= \|\mathbf{x}\|^2 - \sum_{i=1}^n |r_i|^2. \end{aligned}$$

Thus we have  $\|\mathbf{x}\|^2 \geq \sum_i |r_i|^2$ . The second part of the theorem is proven by

$$(\mathbf{x}', \mathbf{x}_j) = (\mathbf{x}, \mathbf{x}_j) - \sum_i r_i^* (\mathbf{x}_i, \mathbf{x}_j) = r_j^* - r_j^* = 0. \quad \clubsuit$$

## 4.2 Hierarchical Structure of Vector Spaces

### 4.2.1 Precise Definitions of Vector Spaces

In this section, we look at the hierarchical structure of abstract vector spaces. We will find that the Hilbert spaces that we have considered form a very limited, special class of general vector spaces under strict conditions. We begin with an exact definition of vector spaces.

#### ♠ Vector spaces:

A vector space  $V$  is a set of elements  $\mathbf{x}$  (called vectors) that satisfy the following sets of axioms:

1.  $V$  is a commutative group under addition:
  - (i)  $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x} \in V$  for any  $\mathbf{x}, \mathbf{y} \in V$  (**closedness**).
  - (ii)  $\mathbf{x} + (\mathbf{y} + \mathbf{z}) = (\mathbf{x} + \mathbf{y}) + \mathbf{z}$  (**associativity**).
  - (iii) There exists an addition identity, the zero vector  $\mathbf{0}$ , for every  $\mathbf{x} \in V$  such that  $\mathbf{x} + \mathbf{0} = \mathbf{x}$ .
  - (iv) There exists an additive inverse  $-\mathbf{x}$  for every  $\mathbf{x} \in V$  such that  $\mathbf{x} + (-\mathbf{x}) = \mathbf{0}$ .
2.  $V$  satisfies the following additional axioms with respect to a number field  $\mathbf{F}$ , whose elements  $\alpha$  are called **scalars**:
  - (i)  $V$  is closed under scalar multiplication:

$$\alpha \mathbf{x} \in V \quad \text{for arbitrary } \mathbf{x} \in V \text{ and } \alpha \in \mathbf{F}.$$

- (ii) Scalar multiplication is distributive with respect to elements of both  $V$  and  $\mathbf{F}$ :

$$\alpha(\mathbf{x} + \mathbf{y}) = \alpha \mathbf{x} + \alpha \mathbf{y}, \quad (\alpha + \beta)\mathbf{x} = \alpha \mathbf{x} + \beta \mathbf{x}.$$

- (iii) Scalar multiplication is associative:  $\alpha(\beta \mathbf{x}) = \beta(\alpha \mathbf{x})$ .
- (iv) Multiplication with the zero scalar  $0 \in \mathbf{F}$  gives the zero vector such that  $0\mathbf{x} = \mathbf{0} \in V$ .
- (v) The unit scalar  $1 \in \mathbf{F}$  has the property that  $1\mathbf{x} = \mathbf{x}$ .

In these definitions,  $\mathbf{F}$  is either the set of real numbers,  $\mathbf{R}$ , or the set of complex numbers,  $\mathbf{C}$ . A vector space over  $\mathbf{R}$  is called a **real vector space**. If  $\mathbf{F} = \mathbf{C}$ , then  $V$  is a **complex vector space**.

### 4.2.2 Metric Space

Once a vector space is endowed with the concept of a *distance* between the elements, say,  $\mathbf{x} \in V$  and  $\mathbf{y} \in V$ , it is called a **metric space**.

♠ **Metric space:**

Assume a vector space  $V$ . A metric space is the pair  $(V, \rho)$  in which the function  $\rho : V \times V \rightarrow \mathbf{R}$ , called the **distance function**, is a single-valued, nonnegative, real function that satisfies:

1.  $\rho(\mathbf{x}, \mathbf{y}) = 0$  if and only if  $\mathbf{x} = \mathbf{y}$ .
2.  $\rho(\mathbf{x}, \mathbf{y}) = \rho(\mathbf{y}, \mathbf{x})$ .
3.  $\rho(\mathbf{x}, \mathbf{y}) \leq \rho(\mathbf{x}, \mathbf{z}) + \rho(\mathbf{z}, \mathbf{y})$  for any  $\mathbf{z} \in V$ .

*Remark.* Strictly speaking, the above is called a metric *vector* space as a subset of more general metric spaces. The latter consists of a pair  $(U, \rho)$ , where  $U$  is a set of points (not necessarily vectors) and  $\rho$  is a distance function. If  $U$  is a vector space  $V$ , then  $(V, \rho)$  is called a **metric vector space**.

*Examples* 1. If we set

$$\rho(\mathbf{x}, \mathbf{y}) = \begin{cases} 0 & \text{if } \mathbf{x} = \mathbf{y}, \\ 1 & \text{if } \mathbf{x} \neq \mathbf{y} \end{cases}$$

for arbitrary  $\mathbf{x}, \mathbf{y} \in V$ , we obtain a metric space.

2. The set of real numbers  $\mathbf{R}$  with the distance function  $\rho(x, y) = |x - y|$  forms a metric space.
3. The set of ordered  $n$ -tuples of real numbers  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  with the distance function

$$\rho(\mathbf{x}, \mathbf{y}) = \left[ \sum_{i=1}^n (x_i - y_i)^2 \right]^{1/2}$$

is a metric space. This is in fact the Euclidean  $n$ -space, denoted by  $\mathbf{R}^n$ .

4. Consider again the set of ordered  $n$ -tuples of real numbers  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  with an alternative distance function:

$$\rho(\mathbf{x}, \mathbf{y}) = \max [|x_i - y_i|; 1 \leq i \leq n].$$

This also serves as a metric space. The validity of Axioms **1–3** mentioned above is obvious.

Comparison between Examples **3** and **4** tells us that the same vector space  $V$  can be metrized in different ways. These two examples call attention to the importance of distinguishing a metric space  $(V, \rho)$  from the vector space  $V$ .

### 4.2.3 Normed Spaces

A metric space is said to be normed if for each element  $\mathbf{x} \in V$  there is a corresponding nonnegative number  $\|\mathbf{x}\|$ , which is called the **norm of  $\mathbf{x}$** .

♠ **Normed space:**

A metric space equipped with a **norm** is called a **normed space**. The norm is defined as a real-valued function (denoted by  $\|\cdot\|$ ) on a vector space  $V$ , which satisfies

1.  $\|\lambda\mathbf{x}\| = |\lambda|\|\mathbf{x}\|$  for all  $\lambda \in \mathbf{F}$  and  $\mathbf{x} \in X$ .
2.  $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ .
3.  $\|\mathbf{x}\| = 0$  if and only if  $\mathbf{x} = \mathbf{0}$ .

Obviously, a normed space is a metric space under the definition of the distance  $\rho(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$ .

*Examples* 1. The space consisting of all  $n$ -tuples of real numbers:  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  in which the norm is defined by

$$\|\mathbf{x}\| = \left( \sum_{i=1}^n x_i^2 \right)^{1/2}$$

is a normed space.

2. The space above can be normed by a more general form:

$$\|\mathbf{x}\| = \left( \sum_{i=1}^n x_i^p \right)^{1/p} \quad (p \geq 1).$$

This norm is referred to as a  **$p$ -norm** of the vector  $\mathbf{x}$ .

3. We further obtain an alternative normed space if we set the norm of the vector  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  equal to the  $\max\{|x_k|; 1 \leq k \leq n\}$ .
4. The collection of all continuous functions defined on the closed interval  $[a, b]$  in which

$$\|f(x)\| \equiv \max\{|f(x)| : x \in [a, b]\}$$

is a normed space.

5. The space consisting of all sequences  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  of real numbers that satisfy the condition  $\lim_{n \rightarrow \infty} x_n = 0$  is a normed space if we set

$$\|\mathbf{x}\| = \max\{|x_k| : 1 \leq k \leq \infty\}.$$

#### 4.2.4 Subspaces of a Normed Space

A class of normed spaces involves the following two subclasses: one endowed with completeness and the other with the inner product. The normed spaces of the former class, i.e., a class of complete normed vector spaces, are called **Banach spaces**.

♠ **Banach space:**

If a normed space is complete, it is called a **Banach space**.

Here, the completeness of a space implies that every Cauchy sequence in the space is convergent. Refer to the arguments in Sect. 4.1.6 for details.

**Remark.** Every finite-dimensional normed space is a Banach space, since it is necessarily complete.

*Examples* **1.** Suppose that a set of infinite-dimensional vectors  $\mathbf{x} = (x_1, x_2, \dots, x_n, \dots)$  satisfies the condition

$$\sum_{i=1}^{\infty} |x_i|^p < \infty, \quad (p \geq 1).$$

Then, this set is a Banach space, called an  $\ell^p$  **space**, under the  $p$ -norm defined by

$$\|\mathbf{x}\|_p = \left( \sum_{i=1}^{\infty} |x_i|^p \right)^{1/p}. \quad (4.11)$$

The proof of its completeness is given in Sect. 4.3.1.

**2.** Assume a set of functions  $f(x)$  expressed by

$$\int_a^b |f(x)|^p dx < \infty.$$

Then, this set constitutes a specific class of Banach spaces, called an  $L^p$  **spaces**, under the  $p$ -norm:

$$\|f\|_p = \left( \int_a^b |f(x)|^p dx \right)^{1/p}. \quad (4.12)$$

Completeness is proved in Sect. 4.3.2.

Now we focus on the counterpart, i.e., a noncompleted normed space endowed with an inner product known as a **pre-Hilbert space**.

♠ **Pre-Hilbert space:**

If a normed space is equipped with an inner product (not necessarily complete), then it is called a **pre-Hilbert space**.

Finally, we are at a point at which we can appreciate the definition of Hilbert spaces. They are defined as the intersection between Banach spaces and pre-Hilbert spaces as stated below

♠ **Hilbert space:**

A complete pre-Hilbert space, i.e., a complete normed space endowed with an inner product is called a **Hilbert space**.

*Examples* The  $\ell^p$  spaces and  $L^p$  spaces with  $p = 2$ , known as the  $\ell^2$  **spaces** and  $L^2$  **spaces**, are Hilbert spaces. The inner product of each space, respectively, is given by

$$(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{\infty} x_i y_i \quad \text{and} \quad (f, g) = \int_a^b f^*(x) g(x) dx. \quad (4.13)$$

*Remark.* Clearly the quantities  $(\mathbf{x}, \mathbf{x})^{1/2}$  and  $(f, f)^{1/2}$ , defined through the inner products (4.13), are special cases of the  $p$ -norm given by (4.11) and (4.12), respectively, with  $p = 2$ . In fact, for the  $\ell^2$  and  $L^2$  spaces, the inner products are defined such that

$$(\mathbf{x}, \mathbf{x}) = \|\mathbf{x}\|^2 \quad \text{and} \quad (f, f) = \|f\|^2.$$

However, for  $\ell^p$  and  $L^p$  spaces with  $p \neq 2$ , we cannot introduce inner products as

$$(\mathbf{x}, \mathbf{x}) = (\|\mathbf{x}\|_p)^p \quad \text{and} \quad (f, f) = (\|f\|_p)^p$$

because unless  $p = 2$  the  $p$ -norm violates the parallelogram law. Accordingly, among the family of  $\ell^p$  and  $L^p$ , only the spaces  $\ell^2$  and  $L^2$  can be Hilbert spaces because they have an inner product.

#### 4.2.5 Basis of a Vector Space: Revisited

For use in Sect. 4.2.6, we briefly review the definition of a **basis** in a *finite-dimensional* vector space and related matters.

♠ **Linearly independent vector:**

A *finite* set of vectors, say,  $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$  is **linearly independent** if and only if

$$\sum_i c_i \mathbf{e}_{i=1}^n = 0 \iff c_i = 0 \text{ for all } i. \quad (4.14)$$

This definition applies to *infinite* sets of vectors  $\mathbf{e}_1, \mathbf{e}_2, \dots$  if the vector space under consideration admits a definition of convergence (see Sect. 4.2.6 for details).

♠ **Basis of a vector space:**

A **basis** of the vector space  $V$  is a set of linearly independent vectors  $\{\mathbf{e}_i\}$  of  $V$  such that every vector  $\mathbf{x}$  of  $V$  can be expressed as

$$\mathbf{x} = \sum_{i=1}^n \alpha_i \mathbf{e}_i. \quad (4.15)$$

Here, the numbers  $\alpha_1, \alpha_2, \dots, \alpha_n$  are **coordinates** of the vector  $\mathbf{x}$  with respect to the basis, and they are uniquely determined owing to the linear independence property.

Therefore, every set of  $n$  linearly independent vectors is a basis in a finite-dimensional vector space spanned by  $n$  vectors. The number  $n$  is called the **dimension** of the vector space. Obviously, an infinite-dimensional vector space does not admit a finite basis, which is why it is called infinite-dimensional.

#### 4.2.6 Orthogonal Bases in Hilbert Spaces

For any vector space (finite- or infinite-dimensional), a set of orthogonal vectors  $\{\mathbf{x}_n\}$  is called an **orthogonal basis** if it is complete. Similarly, a complete orthogonal set of vectors is called an **orthonormal basis** if the norm  $\|\mathbf{x}_n\| = 1$  for all  $n$ . It is convenient to use orthonormal bases in studying Hilbert spaces, since any vector in the space can be decomposed into a linear combination of orthonormal bases. However, when we choose some basis for an infinite-dimensional space, some care must be taken to examine its completeness property; i.e., an infinite sum of vectors in a vector space may or may not be convergent to the identical vector space.

To examine this point, let us consider an infinite set  $\{\mathbf{e}_i\}$  ( $i = 1, 2, \dots$ ) of orthonormal vectors all belonging to a Hilbert space  $V$ . We take any vector  $\mathbf{x} \in V$  and form the set of vectors

$$\mathbf{x}_n = \sum_{i=1}^n c_i \mathbf{e}_i, \quad (4.16)$$

where the complex number  $c_i$  is the inner product of  $\mathbf{e}_i$  and  $\mathbf{x}$  expressed by

$$c_i = (\mathbf{e}_i, \mathbf{x}).$$

For the pair of vectors  $\mathbf{x}$  and  $\mathbf{x}_n$ , the Schwarz inequality (4.2) gives

$$|(\mathbf{x}, \mathbf{x}_n)|^2 \leq \|\mathbf{x}\|^2 \|\mathbf{x}_n\|^2 = \|\mathbf{x}\|^2 \left( \sum_{i=1}^n |c_i|^2 \right). \quad (4.17)$$

On the other hand, taking the inner product of (4.16) with  $\mathbf{x}$  yields

$$(\mathbf{x}, \mathbf{x}_n) = \sum_{i=1}^n c_i (\mathbf{x}, \mathbf{e}_i) = \sum_{i=1}^n |c_i|^2. \quad (4.18)$$

From (4.17) and (4.18), we have

$$\sum_{i=1}^n |c_i|^2 \leq \|\mathbf{x}\|^2.$$

This conclusion is true for arbitrarily large  $n$  and can be stated as shown below.

♠ **Bessel inequality:**

Let  $\{\mathbf{e}_i\}$  ( $i = 1, 2, \dots$ ) be an infinite set of orthonormal vectors in a Hilbert space  $V$ . Then for any  $\mathbf{x} \in V$  with  $c_i = (\mathbf{e}_i, \mathbf{x})$ , we have

$$\sum_{i=1}^{\infty} |c_i|^2 \leq \|\mathbf{x}\|^2,$$

which is known as the **Bessel inequality**.

The Bessel inequality shows that the limiting vector

$$\lim_{n \rightarrow \infty} \sum_{i=1}^n c_i \mathbf{e}_i \equiv \sum_{i=1}^{\infty} c_i \mathbf{e}_i \quad (4.19)$$

has a finite norm, which means that the vector (4.19) is convergent. However, we still do not know whether it converges to  $\mathbf{x}$ . To make such a statement, the set  $\{\mathbf{e}_i\}$  should be equipped with the completeness property defined below.

♠ **Complete orthonormal vectors:**

An infinite set of orthonormal vectors  $\{\mathbf{e}_i\}$  in a Hilbert space  $V$  is called **complete** if the only vector in  $V$  that is orthogonal to all the  $\mathbf{e}_i$  is the zero vector.

The following is an immediate consequence of the above statement.

♠ **Parseval identity:**

Let  $\{\mathbf{e}_i\}$  be an infinite set of orthonormal vectors in a Hilbert space  $V$ . Then for any  $\mathbf{x} \in V$ ,

$$\begin{aligned} & \{\mathbf{e}_i\} \text{ is complete} \\ \iff & \|\mathbf{x}\|^2 = \sum_{i=1}^{\infty} |c_i|^2 \text{ with } c_i = (\mathbf{e}_i, \mathbf{x}). \end{aligned} \quad (4.20)$$

*Proof* Suppose that the set  $\{\mathbf{e}_i\}$  is complete and consider the vector defined by

$$\mathbf{y} = \mathbf{x} - \sum_{i=1}^{\infty} c_i \mathbf{e}_i,$$

where  $\mathbf{x} \in V$  and  $c_i = (\mathbf{e}_i, \mathbf{x})$ . It follows that for any  $\mathbf{e}_j$ ,

$$(\mathbf{e}_j, \mathbf{y}) = (\mathbf{e}_j, \mathbf{x}) - \sum_{i=1}^{\infty} c_i (\mathbf{e}_j, \mathbf{e}_i) = c_j - \sum_{i=1}^{\infty} c_i \delta_{ji} = 0. \quad (4.21)$$

In view of the definition of the completeness of  $\{\mathbf{e}_i\}$ , (4.21) means that  $\mathbf{y}$  is the zero vector. Hence, we have

$$\mathbf{x} = \sum_{i=1}^{\infty} c_i \mathbf{e}_i,$$

which implies

$$\|\mathbf{x}\|^2 = \sum_{i=1}^{\infty} |c_i|^2.$$

We now consider the converse. Suppose  $\mathbf{x}$  to be orthogonal to all the  $\{\mathbf{e}_i\}$ , which means

$$(\mathbf{e}_i, \mathbf{x}) = c_i = 0 \text{ for all } i. \quad (4.22)$$

It follows from (4.20) to (4.22) that  $\|\mathbf{x}\|^2 = 0$ , which in turn gives  $\mathbf{x} = \mathbf{0}$ , because only the zero vector has a zero vector. This completes the proof. ♣

We close this section by providing precise terminology for the basis of a Hilbert space.

♠ **Basis of a Hilbert space:**

A complete orthonormal set  $\{\mathbf{e}_i\}$  ( $i = 1, 2, \dots$ ) in a Hilbert space  $V$  is called a **basis** of  $V$ .

*Remark.*

1. The concept of *completeness* of an orthonormal set of vectors is distinct from the concept of *completeness* of the Hilbert space, but they are mutually related.
2. In order to define **generalized Fourier coefficients**  $c_i = (\mathbf{e}_i, \mathbf{x})$  for  $\mathbf{x} \in V$  (see Sect. 4.3.4), it suffices for the set  $\{\mathbf{e}_i\}$  to be only orthonormal, nor necessarily complete.

## 4.3 Hilbert Spaces of $\ell^2$ and $L^2$

### 4.3.1 Completeness of the $\ell^2$ Spaces

In this subsection, we examine the completeness property of the space  $\ell^2$  on the field  $\mathbf{F}$  (here  $\mathbf{F} = \mathbf{R}$  or  $\mathbf{C}$ ). As already noted, the completeness of a given vector space  $V$  is characterized by the fact that every Cauchy sequence  $(\mathbf{x}_n)$  involved in the space converges to an element  $\mathbf{x} \in V$  such that  $\lim_{n \rightarrow \infty} \|\mathbf{x} - \mathbf{x}_n\| = 0$ . Hence, to prove the completeness of the  $\ell^2$  space, we show in turn that (1) every Cauchy sequence  $(\mathbf{x}_n)$  in the  $\ell^2$  space converges to a limit  $\mathbf{x}$ , and (2) the limit  $\mathbf{x}$  belongs to  $\ell^2$ .

We consider Statement (1). Assume a set of infinite-dimensional vectors

$$\mathbf{x}^{(n)} = (x_1^{(n)}, x_2^{(n)}, \dots),$$

wherein  $x_i^{(n)} \in \mathbf{F}$ , and let the sequence of vectors  $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots\}$  be a Cauchy sequence in the sense of the norm

$$\|\mathbf{x}\| = \left( \sum_{i=1}^{\infty} |x_i|^2 \right)^{1/2} < \infty.$$

Then, for any  $\varepsilon > 0$ , there exists an integer  $N$  such that

$$m, n > N \Rightarrow \left\| \mathbf{x}^{(m)} - \mathbf{x}^{(n)} \right\| = \left( \sum_{i=1}^{\infty} \left| x_i^{(m)} - x_i^{(n)} \right|^2 \right)^{1/2} < \varepsilon. \quad (4.23)$$

This implies that

$$\left| x_i^{(m)} - x_i^{(n)} \right| < \varepsilon \quad (4.24)$$

for every  $i$  and every  $m, n > N$ . Furthermore, since (4.23) is true in the limit  $m \rightarrow \infty$ , we find

$$\left\| \mathbf{x} - \mathbf{x}^{(n)} \right\| < \varepsilon \quad (4.25)$$

for arbitrary  $n > N$ . The inequalities (4.24) and (4.25) mean that  $\mathbf{x}^{(n)}$  converges to the limiting vector expressed by  $\mathbf{x} \equiv (x_1, x_2, \dots)$ , in which the component  $x_i \in \mathbf{F}$  is defined by

$$x_i = \lim_{n \rightarrow \infty} x_i^{(n)}. \quad (4.26)$$

(That the limit (4.26) belongs to  $\mathbf{F}$  is guaranteed by the completeness of  $\mathbf{F}$ .)

The remaining task is to show that the limiting vector  $\mathbf{x}$  belongs to the original space  $\ell^2$ . By the triangle inequality, we have

$$\|\mathbf{x}\| = \left\| \mathbf{x} - \mathbf{x}^{(n)} + \mathbf{x}^{(n)} \right\| \leq \left\| \mathbf{x} - \mathbf{x}^{(n)} \right\| + \left\| \mathbf{x}^{(n)} \right\|.$$

Hence, for every  $n > N$  and for every  $\varepsilon > 0$ , we obtain

$$\|\mathbf{x}\| < \varepsilon + \left\| \mathbf{x}^{(n)} \right\|.$$

As the Cauchy sequence  $(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots)$  is bounded,  $\|\mathbf{x}\|$  cannot be greater than

$$\varepsilon + \limsup_{i \rightarrow \infty} \left\| \mathbf{x}^{(i)} \right\|$$

and is therefore finite. This implies that the limit vector  $\mathbf{x}$  belongs to  $\ell^2(\mathbf{F})$ . Consequently, we have proven that the space  $\ell^2(\mathbf{F})$  is complete.

*Remark.* Among the various kinds of Hilbert spaces, the space  $\ell^2$  has a significant importance in mathematical physics, mainly because it provides the groundwork for the theory of quantum mechanics. In fact, any element  $\mathbf{x}$  of the space  $\ell^2$  satisfying the normalized conditions  $\|\mathbf{x}\| = \sum_{i=1}^{\infty} |x_i|^2 = 1$  works as a possible state vector of quantum systems. In the Heisenberg formulation of quantum mechanics, the infinite-dimensional matrices corresponding to physical observables act on these state vectors.

### 4.3.2 Completeness of the $L^2$ Spaces

We next consider another important class of Hilbert spaces, called  $L^2$  spaces, which are spanned by square-integrable functions  $\{f_n(x)\}$  on a closed interval, say  $[a, b]$ . To prove the completeness of the  $L^2$  space, we show that every

Cauchy sequence  $\{f_n\}$  in the  $L^2$  space converges to a limit function  $f(x)$ , and then verify that the  $f$  belongs to  $L^2$ .

Let  $\{f_1(x), f_2(x), \dots\}$  be a Cauchy sequence in  $L^2$ . Then for any small  $\varepsilon > 0$ , we can find an integer  $N$  such that

$$m, n > N \Rightarrow \|f_n - f_m\| = \sqrt{\int_a^b |f_n(x) - f_m(x)|^2 dx} < \varepsilon.$$

Then, it is always possible to find an integer  $n_1$  such that

$$n > n_1 \Rightarrow \|f_n(x) - f_{n_1}(x)\| < \frac{1}{2}.$$

By mathematical induction, after finding  $n_{k-1} > n_{k-2}$ , we find  $n_k > n_{k-1}$  such that

$$n > n_k \Rightarrow \|f_n(x) - f_{n_k}(x)\| < \left(\frac{1}{2}\right)^k.$$

In this way, we obtain a sequence  $(f_{n_k})$  that is a subsequence such that

$$\|f_{n_{k+1}}(x) - f_{n_k}(x)\| < \left(\frac{1}{2}\right)^k \quad \text{for } k = 1, 2, \dots,$$

or equivalently,

$$\|f_{n_1}\| + \sum_{k=1}^{\infty} \|f_{n_{k+1}} - f_{n_k}\| < \|f_{n_1}\| + \sum_{k=1}^{\infty} \left(\frac{1}{2}\right)^k = \|f_{n_1}\| + 1 \equiv A,$$

where  $A$  is a finite constant. Let

$$g_k = |f_{n_1}| + |f_{n_2} - f_{n_1}| + \dots + |f_{n_{k+1}} - f_{n_k}| \quad (k = 1, 2, \dots).$$

Then, by the Minkowski inequality, we have

$$\begin{aligned} \int_a^b [g_k(x)]^2 dx &= \int_a^b [|f_{n_1}| + |f_{n_2} - f_{n_1}| + \dots + |f_{n_{k+1}} - f_{n_k}|]^2 dx \\ &\leq \left( \|f_{n_1}\| + \sum_{i=1}^k \|f_{n_{i+1}} - f_{n_i}\| \right)^2 \leq A^2 < \infty. \end{aligned} \quad (4.27)$$

Let  $g(x) = \lim g_k(x)$ . Then  $[g(x)]^2 = \lim [g_k(x)]^2$ , and

$$\int_a^b [g(x)]^2 dx = \int_a^b \lim_{k \rightarrow \infty} [g_k(x)]^2 dx = \lim_{k \rightarrow \infty} \int_a^b [g_k(x)]^2 dx. \quad (4.28)$$

[See the remark below for the interchangeability of the limit and integral signs in (4.28).] It follows from (4.27) and (4.28) that

$$\int_a^b [g(x)]^2 dx < \infty,$$

or equivalently,

$$\int_a^b \left( |f_{n_1}| + \sum_{k=1}^{\infty} |f_{n_{k+1}} - f_{n_k}| \right)^2 dx < \infty.$$

This implies that the infinite sum

$$\|f_{n_1}\| + \sum_{k=1}^{\infty} \|f_{n_{k+1}} - f_{n_k}\| \quad (4.29)$$

converges to a function, denoted by  $f \in L^2$ , in the sense of the norm in  $L^2$ .

We next show that the limit function  $f(x)$  expressed by (4.29) is an element of  $L^2$  such as

$$\|f_n(x) - f(x)\| \rightarrow 0 \quad (n \rightarrow \infty). \quad (4.30)$$

We first note that

$$f(x) - f_{n_j}(x) = \sum_{k=j}^{\infty} [f_{n_{k+1}} - f_{n_k}(x)].$$

It follows that

$$\|f - f_{n_j}\| \leq \sum_{k=j}^{\infty} \|f_{n_{k+1}} - f_{n_k}\| < \sum_{k=j}^{\infty} \left(\frac{1}{2}\right)^k = \frac{1}{2^{j-1}},$$

so we have

$$\lim_{j \rightarrow \infty} \|f - f_{n_j}\| = 0.$$

Observe that

$$\|f_n - f\| \leq \|f_n - f_{n_k}\| + \|f_{n_k} - f\|,$$

where  $\|f_n - f_{n_k}\| \rightarrow 0$  as  $n \rightarrow \infty$  and  $k \rightarrow \infty$ ; thus

$$\lim_{n \rightarrow \infty} \|f_n - f\| = 0,$$

which shows that the Cauchy sequence  $(f_n)$  converges to  $f \in L^2$ .

**Remark.** The interchangeability of limit and integral signs in (4.28) is justified by the following three facts:

- (i) The sequence  $([g_k(x)]^2)$  is a sequence of square-integrable functions in  $[a, b]$ ,
- (ii)  $[g_k(x)]^2 \geq 0$  for all  $k$ , and
- (iii) The integral  $\int_a^b [g_k]^2 dx$  for each  $k$  has a common bound  $A^2$  as shown in (4.27). The proof of this point is based on the theory of the Lebesgue integral, which we discuss in Chap. 6.

### 4.3.3 Mean Convergence

Before proceeding, comments on a new class of convergence that is relevant to the argument on the completeness of the  $L^2$  space are in place. Observe that the expression (4.30) is rephrased in the following sentence: For any small  $\varepsilon > 0$ , it is possible to find  $N$  such that

$$n > N \Rightarrow \|f(x) - f_n(x)\| < \varepsilon. \quad (4.31)$$

Hence, we can say that the infinite sequence  $(f_n)$  converges to  $f(x)$  in the norm of the  $L^2$  space. Convergence of the type (4.31) is called **the convergence in the mean** or **the mean convergence**, which is inherently different from the uniform convergence and the pointwise convergence. The point is the fact that in the mean convergence, the quantitative deviation between  $f_n(x)$  and  $f(x)$  is measured not by the difference  $f(x) - f_n(x)$ , but by the norm in the  $L^2$  space based on the integration procedure:

$$\|f(x) - f_n(x)\| = \left[ \int_a^b |f(x) - f_n(x)|^2 dx \right]^{1/2}.$$

Hence, when  $f(x)$  is convergent in the mean to  $f_n(x)$  on the interval  $[a, b]$ , there may exist a finite number of **isolated points** such that  $f(x) \neq f_n(x)$ . Obviously, this situation is not allowed in cases of uniform or pointwise convergence.

### 4.3.4 Generalized Fourier Coefficients

Having clarified the completeness property of the two specific Hilbert spaces,  $\ell^2$  and  $L^2$ , we introduce two important concepts: **generalized Fourier coefficients** and **generalized Fourier series**. We shall see that they play a crucial role in revealing the close relationship between the two distinct Hilbert spaces  $\ell^2$  and  $L^2$ .

#### ♠ Generalized Fourier coefficients:

Suppose that a set of square-integrable functions  $\{\phi_i\}$  is orthonormal (not necessarily complete) in the norm of the  $L^2$  space. Then, the numbers

$$c_k = (f, \phi_k) \quad (4.32)$$

are called the **Fourier coefficients** of the function  $f \in L^2$  relative to the orthonormal set  $\{\phi_i\}$ , and the series

$$\sum_{k=1}^{\infty} c_k \phi_k \quad (4.33)$$

is called the **Fourier series** of  $f$  with respect to the set  $\{\phi_i\}$ .

*Remark.*

1. In general, the Fourier series shown in (4.33) may or may not be convergent; its convergence property is determined by the features of the functions  $f$  and the associated orthonormal set of functions  $\{\phi_k\}$ .
2. Some readers may be familiar with the Fourier series associated with trigonometric functions or imaginary exponentials. Notably, however, the concepts of Fourier series and Fourier coefficients introduced above are more general concepts than those associated with trigonometric series.

The importance of the Fourier coefficients (4.32) becomes apparent when we see that they consist of the  $\ell^2$  space. In fact, since  $c_k$  is the inner product of  $f$  and  $\phi_k$ , it yields the **Bessel inequality** in terms of  $c_k$  and  $f$ :

$$\sum_{k=1}^{\infty} |c_k|^2 \leq \|f\|. \quad (4.34)$$

From the hypothesis of  $f \in L^2$ , the norm  $\|f\|$  remains finite. Hence, the inequality (4.34) ensures the convergence of the infinite series  $\sum_{k=1}^{\infty} |c_k|^2$ , which consists of the Fourier coefficients defined by (4.32). This convergence means that the sequence of Fourier coefficients  $\{c_k\}$  is an element of the space  $\ell^2$ , whichever orthonormal set of functions  $\phi_k(x)$  we choose. In this context, the two elements  $f \in L^2$  and  $\mathbf{c} = (c_1, c_2, \dots) \in \ell^2$  are connected via the Fourier coefficient (4.32).

### 4.3.5 Riesz–Fisher Theorem

Recall that every Fourier coefficient satisfies the Bessel inequality (4.34). Hence, in order for a given set of complex numbers  $(c_i)$  to constitute the Fourier coefficients of a function  $f \in L^2$ , it is necessary that the series

$$\sum_{k=1}^{\infty} |c_k|^2$$

converge. As a matter of fact, this condition is not only necessary, but also sufficient as stated in the theorem below.

#### ♠ Riesz–Fisher theorem:

Given any set of complex numbers  $(c_i)$  such that

$$\sum_{k=1}^{\infty} |c_k|^2 < \infty, \quad (4.35)$$

there exists a function  $f \in L^2$  such that

$$c_k = (f, \phi_k) \quad \text{and} \quad \sum_{k=1}^{\infty} |c_k|^2 = \|f\|^2, \quad (4.36)$$

where  $\{\phi_i\}$  is a complete orthonormal set.

*Proof* Set linear combinations of  $\phi_k(x)$  as

$$f_n(x) = \sum_{k=1}^n c_k \phi_k(x), \quad (4.37)$$

where the  $c_k$  are arbitrary complex numbers satisfying condition (4.35). Then, for a given integer  $p \geq 1$ , we obtain

$$\|f_{n+p} - f_n\|^2 = \|c_{n+1}\phi_{n+1} + \cdots + c_{n+p}\phi_{n+p}\|^2 = \sum_{k=n+1}^{n+p} |c_k|^2. \quad (4.38)$$

Let  $p = 1$  and  $n \rightarrow \infty$ . Then, from condition (4.35), we have

$$\|f_{n+1} - f_n\| = |c_{n+1}| \rightarrow 0 \quad (n \rightarrow \infty).$$

This tells us that the infinite sequence  $\{f_n\}$  defined by (4.37) associated with a given set of complex numbers  $\{c_i\}$  always converges in the mean to a function  $f \in L^2$ .

Our remaining task is to show that this limit function  $f(x)$  satisfies condition (4.36), so we consider the inner product

$$(f, \phi_i) = (f_n, \phi_i) + (f - f_n, \phi_i), \quad (4.39)$$

where we assume  $n \geq i$ . It follows from (4.37) that the first term on the right-hand side is equal to  $c_i$ . The second term vanishes as  $n \rightarrow \infty$ , since

$$|(f - f_n, \phi_i)| \leq \|f - f_n\| \cdot \|\phi_i\| \rightarrow 0 \quad (n \rightarrow \infty),$$

where we used the mean convergence of  $\{f_n\}$  to  $f$ . In addition, the left-hand side of (4.39) is independent of  $n$ . Hence, taking the limit  $n \rightarrow \infty$  on both sides of (4.39), we obtain

$$(f, \phi_i) = c_i, \quad (4.40)$$

which means that  $c_i$  is the Fourier coefficient of  $f$  relative to  $\phi_i$ . From our assumption, the set  $\{\phi_i\}$  is complete and orthonormal. Hence, the Fourier coefficients (4.40) satisfy the **Parseval identity**:

$$\sum_{k=1}^{\infty} |c_k|^2 = \|f\|^2. \quad (4.41)$$

The results (4.40) and (4.41) are identical to condition (4.36), thus proving the theorem. ♣

### 4.3.6 Isomorphism between $\ell^2$ and $L^2$

The Riesz–Fisher theorem results immediately in the isomorphism between the Hilbert spaces  $L^2$  and  $\ell^2$ . An isomorphism is a one-to-one correspondence that preserves the entire algebraic structure. For instance, two vector spaces  $U$  and  $V$  (over the same number field) are isomorphic if there exists a one-to-one correspondence between the vectors  $\mathbf{x}_i$  in  $U$  and  $\mathbf{y}_i$  in  $V$ , say  $\mathbf{y}_i = f(\mathbf{x}_i)$ , such that

$$f(\alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2) = \alpha_1 f(\mathbf{x}_1) + \alpha_2 f(\mathbf{x}_2).$$

The isomorphism between  $L^2$  and  $\ell^2$  is closely related to the theory of quantum mechanics, which originally consisted of two distinct theories: Heisenberg’s matrix mechanics, based on infinite-dimensional vectors, and Schrödinger’s wave mechanics, based on square-integrable functions. From the mathematical point of view, the difference between the two theories reduces to the fact that the former uses the space  $\ell^2$ , whereas the latter uses the space  $L^2$ . Hence, the isomorphism between the two spaces verifies the equivalence of the two theories describing the nature of quantum mechanics.

Let us prove the above point. Choose an arbitrary complete orthonormal set  $\{\phi_n\}$  in  $L^2$  and assign to each function  $f \in L^2$  the sequence  $(c_1, c_2, \dots, c_n, \dots)$  of its Fourier coefficients with respect to this set. Since

$$\sum_{k=1}^{\infty} |c_k|^2 = \|f\|^2 < \infty,$$

the sequence  $(c_1, c_2, \dots, c_n, \dots)$  is an element of  $\ell^2$ . Conversely, in view of the Riesz–Fisher theorem, for every element  $(c_1, c_2, \dots, c_n, \dots)$  of  $\ell^2$  there is a function  $f(x) \in L^2$  whose Fourier coefficients are  $c_1, c_2, \dots, c_n, \dots$ . This correspondence between the elements of  $L^2$  and  $\ell^2$  is one-to-one. Furthermore, if

$$f(x) \longleftrightarrow (c_1, c_2, \dots, c_n, \dots)$$

and

$$g(x) \longleftrightarrow (d_1, d_2, \dots, d_n, \dots),$$

then

$$f(x) + g(x) \longleftrightarrow (c_1 + d_1, \dots, c_n + d_n, \dots)$$

and

$$kf(x) \longleftrightarrow (kc_1, kc_2, \dots, kc_n, \dots),$$

which readily follows from the definition of Fourier coefficients (the reader should prove it). That is, addition and multiplication by scalars are preserved by the correspondence. Furthermore, in view of Parseval’s identity, it follows that

$$(f, g) = \sum_{i=1}^{\infty} c_i^* d_i. \quad (4.42)$$

All of these facts ensure the isomorphism between the spaces  $L^2$  and  $\ell^2$ , i.e., the one-to-one correspondence between the elements of  $L^2$  and  $\ell^2$  that preserves the algebraic structures of the space. In this context, we may say that every element  $\{c_i\}$  in an  $\ell^2$  space serves as a *coordinate system* of the  $L^2$  space, and vice versa.

## Exercises

1. Prove the inequality  $\sum_{k=1}^{\infty} |c_k|^2 \leq \|f\|^2$  given in (4.34).

**Solution:** Suppose a partial sum  $S_n(x) = \sum_{k=1}^n \alpha_k \phi_k(x)$ , where  $\alpha_k$  is a certain number (real or complex). Since the set  $\{\phi_i\}$  is orthonormal,

$$\begin{aligned} \|f(x) - S_n(x)\|^2 &= \left\| f - \sum_{j=1}^n \alpha_j \phi_j(x), f - \sum_{k=1}^n \alpha_k \phi_k(x) \right\|^2 \\ &= \|f\|^2 - \sum_{k=1}^n |c_k|^2 + \sum_{k=1}^n (\alpha_k - c_k)^2. \end{aligned} \quad (4.43)$$

The minimum of (4.43) is assumed if  $\alpha_k = c_k$ . In that case, the equation (4.43) reads

$$\|f(x) - \sum_{k=1}^n c_k \phi_k(x)\|^2 = \|f\|^2 - \sum_{k=1}^n |c_k|^2,$$

which implies  $\sum_{k=1}^n |c_k|^2 \leq \|f\|^2$ . Since the right-hand side is independent of  $n$ , the value of  $n$  can be taken arbitrarily large. Hence, by taking the limit  $n \rightarrow \infty$ , we attain the desired result:  $\sum_{k=1}^{\infty} |c_k|^2 \leq \|f\|^2$ . ♣

2. Verify the equation  $(f, g) = \sum_{i=1}^{\infty} c_i^* d_i$  given in (4.42).

**Solution:** This equality is verified because of the relations  $(f, f) = \sum_{i=1}^{\infty} |c_i|^2$  and  $(g, g) = \sum_{i=1}^{\infty} |d_i|^2$ , and their consequences:

$$\begin{aligned} (f + g, f + g) &= (f, f) + 2(f, g) + (g, g) = \sum_{i=1}^{\infty} |c_i + d_i|^2 \\ &= \sum_{i=1}^{\infty} |c_i|^2 + 2 \sum_{i=1}^{\infty} c_i^* d_i + \sum_{i=1}^{\infty} |d_i|^2. \quad \clubsuit \end{aligned}$$



## Orthonormal Polynomials

**Abstract** The theory of Hilbert spaces we dealt with in Chap. 4 can be used to construct a number of polynomial functions that are orthonormal and complete in the sense of the  $L^p$  space. In this chapter we present three important approaches for the construction of orthonormal polynomials, based, respectively, on the Weierstrass theorem (Sect. 5.1.1), the Rodrigues formula (Sect. 5.2.1), and generating functions (Sect. 5.2.7). We shall find that various orthonormal polynomials relevant to mathematical physics can be effectively classified by adopting these methods.

### 5.1 Polynomial Approximations

#### 5.1.1 Weierstrass Theorem

There are a number of special polynomials that play a significant role in various aspects of mathematical physics: Legendre, Laguerre, Hermite, and Chebyshev polynomials are well known. For instance, Legendre and Laguerre polynomial expansions are often used to solve second-order differential equations having spherical symmetry. The point is that many of these special polynomials form a **complete orthonormal set of polynomials**; the origin of their orthonormality and completeness can be accounted for in terms of the theory of the Hilbert space  $L^2$ . Owing to completeness, these special polynomials enable us to produce polynomial approximations of fairly arbitrary functions with desired accuracy, which serves as a useful device in manipulating square-integrable functions.

The validity of polynomial approximations is based on the famous **Weierstrass approximation theorem**, which states that from the set of powers of a real variable  $x$  one can construct a sequence of polynomials that converges uniformly to any continuous function within a finite interval  $[a, b]$ . From this result, we shall see that it is possible to find various kinds of complete orthonormal sets of polynomials on any interval  $[a, b]$ .

In what follows, for simplicity we focus on polynomial approximations only of real-valued functions of a real variable. In the case of a complex-valued

function, the separate validity of the theorem for each of its real and imaginary parts ensures the validity of the theorem.

♠ **Weierstrass approximation theorem:**

If a function  $f(x)$  is continuous on the closed interval  $[a, b]$ , there exists a polynomial such as

$$G_n(x) = \sum_{k=0}^n c_k x^k \quad (5.1)$$

that converges uniformly to  $f(x)$  on  $[a, b]$ .

The proof will be given in Appendix C. Several remarks on this theorem are given below.

- In the polynomial approximation based on (5.1), the values of coefficients  $c_m^{(n)}$  depend on  $n$  for fixed  $m$ . Thus, in order to improve the accuracy of the approximation by going to polynomials of higher degree, the earlier coefficients must change. For instance, when the approximating polynomial (5.1) is replaced by

$$G_{n+1}(x) = \sum_{k=0}^{n+1} d_k x^k,$$

we have in general

$$c_k \neq d_k \text{ for all } k(\leq n).$$

This situation is in contrast to the case of our familiar **Taylor series expansions**, in which the earlier coefficients remain unchanged.

- The Weierstrass theorem requires only that the continuity of functions be approximated. This condition is much weaker than Taylor's theorem for expansion in power series, in which the derivatives of all orders must exist (i.e., it must be **analytic**; see Sect. 7.1.2 for the definition of **analytic functions**). Furthermore, the former theorem can apply to polynomial approximations outside the radius of convergence (see Sect. 7.4.1) of a Taylor series.
- The Weierstrass theorem may be extended to functions of more than one variable. By a straightforward generalization of the proof (see Appendix C), it can be shown that if a function  $f(x_1, x_2, \dots, x_m)$  is continuous in each variable  $x_i$  located within  $[a_i, b_i]$  ( $i = 1, 2, \dots, m$ ), it may be approximated uniformly by the polynomials

$$G_n(x_1, x_2, \dots, x_m) = \sum_{k_1=0}^{N_1} \sum_{k_2=0}^{N_2} \cdots \sum_{k_m=0}^{N_m} c_{k_1 k_2 \dots k_m} x^{k_1} x^{k_2} \cdots x^{k_m}.$$

The special cases of  $m = 2$  and  $m = 3$  are considered in Sects. 5.1.4 and 5.1.5.

### 5.1.2 Existence of Complete Orthonormal sets of Polynomials

It must be emphasized that the Weierstrass theorem requires that the set of polynomials  $\{G_n\}$  be neither orthogonal nor complete. Nevertheless, the theorem ensures indirectly the existence of a variety of *complete orthonormal* sets of polynomials in terms of the  $L^2$  space. The proof of their existence is based on the **Gram-Schmidt orthogonalization method** shown below.

♠ **Gram-Schmidt orthogonalization method:**

Given any set of linearly independent functions  $\{\varphi_i\}$  normalizable on a closed interval, it is possible to construct an orthonormal set of functions  $\{Q_i\}$  through the recursion formula

$$Q_i(x) = \frac{u_i(x)}{\|u_i(x)\|}, (i = 1, 2, \dots),$$

with the definitions:

$$u_1(x) = \varphi_1(x), \quad u_i(x) = \varphi_i(x) - \sum_{k=1}^i (u_k, \varphi_{i+1}) u_k(x).$$

Here,  $(u_k, \varphi_{i+1})$  means the inner product in terms of the  $L^2$  space. Let us apply the Gram-Schmidt orthogonalization process to a set of powers  $\{x^n\}$  that is linearly independent. We then obtain an orthonormal set  $\{Q_i\}$  given by

$$Q_i(x) = \sum_{m=0}^i b'_m{}^{(i)} x^m. \quad (5.2)$$

Owing to the orthogonality of the set  $\{Q_i\}$ , the original functions  $x^m$  are expressed conversely by linear combinations of  $\{Q_i\}$  such as

$$x^m = \sum_{i=0}^m b_i^{(m)} Q_i(x). \quad (5.3)$$

Substituting (5.3) into (5.1), we obtain

$$G_n(x) = \sum_{m=0}^n a_m^{(n)} \sum_{i=0}^m b_i^{(m)} Q_i(x). \quad (5.4)$$

The superscripts  $(n)$  and  $(m)$  attached to the coefficients  $a_m^{(n)}$  and  $b_i^{(m)}$ , respectively, remind us that the values of the terms contained in the finite sequences,

$$\left\{a_0^{(n)}, a_1^{(n)}, \dots, a_n^{(n)}\right\} \quad \text{and} \quad \left\{b_0^{(m)}, b_1^{(m)}, \dots, b_m^{(m)}\right\},$$

depend on  $n$  or  $m$ : as  $n$  (or  $m$ ) increases, all the earlier terms in the sequence must be altered.

Now let us show the completeness of the orthonormal set  $\{Q_n(x)\}$  given by (5.2), which was deduced from the orthogonalization process; this is achieved by proving that Parseval's identity,

$$\sum_{n=1}^{\infty} |(f, Q_n)|^2 = \|f\|^2,$$

holds for any  $f \in L^2$ , or equivalently, by proving that

$$(f, Q_n) = 0 \quad \text{for all } n \iff \|f\| = 0. \quad (5.5)$$

The sentence " $\|f\| = 0$  implies  $(f, Q_n) = 0$  for all  $n$ " immediately follows from the **Bessel inequality**,

$$\sum_{k=0}^{\infty} |(f, Q_k)|^2 \leq \|f\|^2.$$

To prove the converse, we note that if  $(f, Q_n) = 0$  for all  $n$ , we have

$$(f, G_n) = 0 \quad \text{for all } n, \quad (5.6)$$

since the  $G_n$  are linear combinations of the  $Q_n$ . In addition, we recall that the Weierstrass theorem guarantees the uniform convergence of the sequence  $(G_n)$  to  $f$ . Since uniform convergence implies a mean convergence, we obtain

$$\|f - G_n\| \rightarrow 0. \quad (5.7)$$

From (5.6) and (5.7), it follows that

$$\|f - G_n\|^2 \equiv (f - G_n, f - G_n) = \|f\|^2 + \|G_n\|^2 \rightarrow 0,$$

which implies that  $\|f\| = 0$  as well as  $\|G_n\|^2 \rightarrow 0$ . (This is because  $\|f\|^2$  is independent of  $n$  and  $\|G_n\|^2$  is nonnegative for all  $n$ .) As summarized, we attain the desired conclusion (5.5), which indicates that the orthonormal set  $\{Q_n\}$  is complete in terms of the  $L^2$  space.

The completeness of the set  $\{Q_i\}$  means that there exists a set of constants  $\{c_i\}$  such that any function  $g \in L^2$  can be approximated in the mean by the following sequence of partial sums:

$$g_n(x) = \sum_{i=0}^n c_i Q_i(x). \quad (5.8)$$

The reader should appreciate a crucial difference between (5.1) and (5.8). In the latter, the  $c_i$  are independent of  $n$  in contrast to the case of (5.1). Thus as we extend the sum to infinity, the approximation improves without changing the earlier  $c_i$ . Therefore, we may say that there exists an infinite series

$$\lim_{n \rightarrow \infty} g_n(x) = \sum_{i=0}^{\infty} c_i Q_i(x)$$

that converges to  $g$  in the mean. The expansion coefficients  $c_i = (g, Q_i)$  in the infinite series are the **Fourier coefficients** we introduced in Sect. 4.3.4.

### 5.1.3 Legendre Polynomials

The previous discussion revealed that the orthonormal set  $\{Q_i\}$  constructed from the orthogonalization process based on the set of powers  $\{x^m\}$  is complete, so that the linear combination  $\sum_{i=0}^n c_i Q_i$  converges in the mean to  $f \in L^2$ . Let us employ this result to find an explicit function form of a complete orthonormal set of functions  $\{P_n\}$  defined on the interval  $[-1, 1]$ . The first member of such a complete orthogonal set is  $P_0(x) = 1$  (For convenience, the normalization constant is omitted temporarily). Using the Gram-Schmidt orthogonalization process, we have

$$P_1(x) = \frac{x - (x, P_0) P_0}{\|x - (x, P_0) P_0\|} = x,$$

$$P_2(x) = \frac{x^2 - (x^2, P_0) P_0 - (x^2, P_1) P_1}{\|x^2 - (x^2, P_0) P_0 - (x^2, P_1) P_1\|} = \frac{1}{2} (3x^2 - 1),$$

where we use the notation

$$(x^m, P_n) = \int_{-1}^1 x^m P_n(x) dx.$$

Successive procedures give

$$P_3(x) = \frac{1}{2} (5x^3 - 3x), \quad P_4(x) = \frac{1}{8} (35x^4 - 30x^2 + 3),$$

$$P_5(x) = \frac{1}{8} (63x^5 - 70x^3 + 15x), \dots$$

Eventually, we obtain the complete orthonormal set of polynomials  $\{P_n\}$  known as the **Legendre polynomial**. The  $x$  dependence of each function is plotted in Fig. 5.1. Note that  $P_n(x)$  has exactly  $n - 1$  distinct zeros in the open interval  $[-1, 1]$ .

A general formula for  $P_n(x)$  is given by

$$P_n(x) = \frac{1}{2^n} \sum_{k=0}^{[n/2]} (-1)^k \frac{(2n-2k)!}{k! (n-k)! (n-2k)!} x^{n-2k}, \quad (5.9)$$

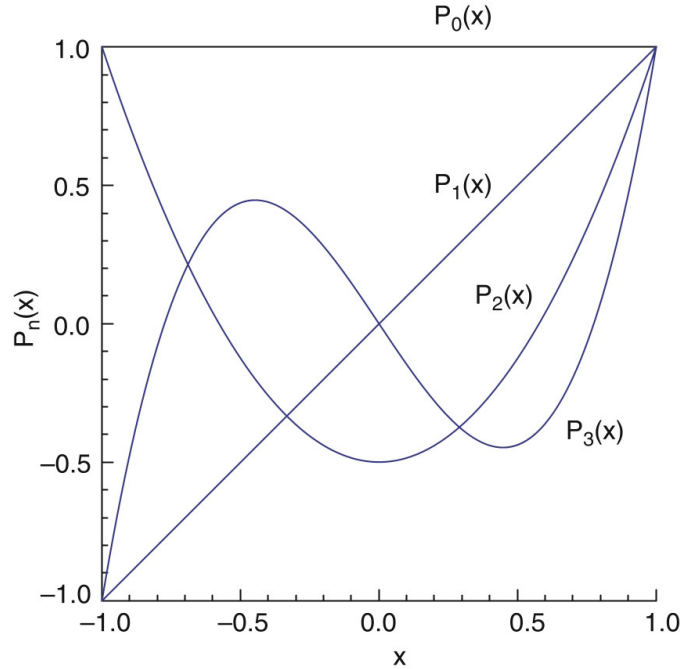
where we used the **Gauss notation**:

$$\left[ \frac{n}{2} \right] = \begin{cases} \frac{n}{2} & \text{if } n \text{ is even,} \\ \frac{n-1}{2} & \text{if } n \text{ is odd.} \end{cases}$$

Equation (5.9) is rewritten in a simpler form as

$$\begin{aligned} P_n(x) &= \frac{1}{2^n} \sum_{k=0}^{[n/2]} \frac{(-1)^k}{k! (n-k)!} \frac{d^n}{dx^n} x^{2n-2k} \\ &= \frac{1}{2^n n!} \frac{d^n}{dx^n} \sum_{k=0}^n \frac{(-1)^k n!}{k! (n-k)!} x^{2n-2k} \\ &= \frac{1}{2^n n!} \frac{d^n}{dx^n} (x^2 - 1)^n. \end{aligned} \quad (5.10)$$

The last line is known as the **Rodrigues formula** for Legendre polynomials. This is a special form of the more general Rodrigues formula that is applicable to any orthonormal polynomial function. The derivations of (5.9) and



**Fig. 5.1.** Profiles of the first three terms of the Legendre polynomial  $P_n(x)$

(5.10), as well as that of general Rodrigues formula are given in Sects. 5.2.1 and 5.2.2.

The orthogonality of the Legendre polynomials follows from the Rodrigues formula (5.10). To see this, we denote  $d^n/dx^n$  by  $d_n$ , and assume that  $n \geq m$ . Dropping constant factors, we have

$$\begin{aligned} \int_{-1}^1 P_n(x)P_m(x)dx &= \int_{-1}^1 [d_n(x^2-1)^n] [d_m(x^2-1)^m] dx \\ &= [d_{n-1}(x^2-1)^n] [d_m(x^2-1)^m] \Big|_{-1}^1 \\ &\quad - \int_{-1}^1 [d_{n-1}(x^2-1)^n] [d_{m+1}(x^2-1)^m] dx, \end{aligned} \quad (5.11)$$

where we employed integration by parts. Since

$$d_{n-1}(x^2-1)^n = (x^2-1) \times (\text{a polynomial}),$$

the first term in the last line of (5.11) vanishes upon putting in the limits  $\pm 1$ , leaving the second term alone. Therefore, after  $n$  partial integrations, we have

$$\int_{-1}^1 P_m(x)P_n(x)dx = (-1)^n \int_{-1}^1 (x^2-1)^n d_{m+n}(x^2-1)^m dx.$$

Now, if  $n > m$ , then  $n+m > 2m$  so that  $d_{n+m}(x^2-1)^m = 0$ . Therefore,

$$\int_{-1}^1 P_n(x)P_m(x)dx = 0 \quad \text{for } m \neq n.$$

If  $m = n$ , then we have

$$\int_{-1}^1 P_n(x)^2 dx = \frac{(-1)^n}{2^{2n}(n!)^2} \int_{-1}^1 (x^2-1)^n d_{2n}(x^2-1)^n dx, \quad (5.12)$$

where a normalization constant is explicitly attached. Since  $(x^2-1)^n$  is a polynomial of degree  $2n$ , its  $(2n)$ th derivative is just  $(2n)!$ . Hence, the integral (5.12) reads

$$\int_{-1}^1 P_n(x)^2 dx = \frac{(2n)! \cdot (-1)^n}{2^{2n}(n!)^2} \int_{-1}^1 (x^2-1)^n dx = \frac{2}{2n+1}. \quad (5.13)$$

As summarized, the orthogonal property of Legendre polynomial functions is given by

$$\int_{-1}^1 P_m(x)P_n(x)dx = \begin{cases} 0 & (m \neq n) \\ \frac{2}{2n+1} & (m = n). \end{cases}$$

*Remark.* Equation (5.13) follows from the identity

$$\begin{aligned}\int_{-1}^1 (1-x^2)^n dx &= 2^{2n+1} \int_0^1 t^n (1-t)^n dt = 2^{2n+1} B(n+1, n+1) \\ &= 2^{2n+1} \frac{\Gamma(n+1)^2}{\Gamma(2n+2)} = 2^{2n+1} \frac{(n!)^2}{(2n+1)!}.\end{aligned}$$

Here, we have changed the variable by setting  $x = 2t - 1$  to obtain the **beta function**  $B(x, y)$  and the **gamma function**  $\Gamma(x)$  defined, respectively, by

$$\begin{aligned}B(x, y) &= \int_0^1 t^{x-1} (1-t)^{y-1} dt = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}, \\ \Gamma(x) &= \int_0^\infty e^{-t} t^{x-1} dt.\end{aligned}$$

#### 5.1.4 Fourier Series

We next consider the application of the Weierstrass theorem to functions of two variables. Through earlier discussions, we have the proof of the completeness properties of the set of trigonometric functions  $\sin n\theta$  and  $\cos n\theta$  ( $n = 0, 1, \dots, \infty$ ).

The Weierstrass theorem tells us that any function  $g(x, y)$  that is continuous in both variables on finite closed intervals may be approximated uniformly by the sequence of functions

$$g_N(x, y) = \sum_{n,m=0}^N a_{nm}^{(N)} x^n y^m. \quad (5.14)$$

Employ polar coordinates and restrict the domain of definition to the unit circle  $x = \cos \theta$  and  $y = \sin \theta$  to find

$$g_N(\cos \theta, \sin \theta) = f_N(\theta) = \sum_{n,m=0}^N a_{nm}^{(N)} \cos^n \theta \sin^m \theta. \quad (5.15)$$

Clearly,  $f_N(\theta)$  should be periodic with periodicity  $2\pi$ . Using **Euler's equation**,

$$e^{i\theta} = \cos \theta + i \sin \theta,$$

we obtain expressions for the  $n$ th powers of  $\sin \theta$  and  $\cos \theta$ :

$$\cos^n \theta = \left[ \frac{1}{2} (e^{i\theta} + e^{-i\theta}) \right]^n, \quad \sin^n \theta = \left[ \frac{1}{2i} (e^{i\theta} - e^{-i\theta}) \right]^n.$$

We then rewrite (5.15) in the form

$$f_M(x) = \sum_{n=-M}^M \frac{c_n^{(M)}}{(2\pi)^{1/2}} e^{inx} \quad \text{with } M = 2N. \quad (5.16)$$

where we have inserted the factor  $(2\pi)^{1/2}$  for later convenience and have replaced the variable  $\theta$  by  $x$  to emphasize the generality of the result.

The superscript  $M$  attached to  $c_n^{(M)}$  in (5.16) suggests the possibility that the values of  $c_n^{(M)}$  are dependent of  $M$ . However, this is not the case. In fact, the values of the coefficients  $c_n$  are determined independently of  $M$  owing to the completeness of the orthonormal set of functions

$$F_n(x) = \frac{e^{inx}}{(2\pi)^{1/2}}, \quad n = 0, \pm 1, \dots,$$

defined on the interval  $[-\pi, \pi]$ . The completeness of the set  $\{F_n\}$  allows us to approximate an arbitrary function  $f$  in the mean by an infinite series of the  $F_n$ , and we write

$$f(x) \simeq \sum_{n=-\infty}^{\infty} c_n F_n(x) = \sum_{n=-\infty}^{\infty} \frac{c_n}{(2\pi)^{1/2}} e^{inx}, \quad (5.17)$$

where the expansion coefficients are given by

$$c_n = (F_n, f) = \frac{1}{(2\pi)^{1/2}} \int_{-\pi}^{\pi} f(x) e^{-inx} dx. \quad (5.18)$$

The series (5.17) with the coefficients (5.18) is known as the **trigonometric Fourier series**. The completeness of the set  $\{F_n\}$  can be verified in a discussion similar to that in Sect. 5.1.2.

### 5.1.5 Spherical Harmonic Functions

We have derived the sets of Legendre polynomials and trigonometric functions from the Weierstrass approximation theorem in one and two variables, respectively. We now derive the set of spherical harmonics from a three-variable generalization. It tells us that a function  $g$  of  $x, y, z$  (i.e.,  $\mathbf{r}$ ) can be approximated uniformly by a sequence of partial sums given by

$$g_M(\mathbf{r}) = \sum_{j,k,n=0}^M a_{jkn}^{(M)} x^j y^k z^n. \quad (5.19)$$

We may also use an alternative coordinate system such as

$$\begin{aligned} u &\equiv x + iy = r \sin \theta e^{i\phi}, \\ v &\equiv x - iy = r \sin \theta e^{-i\phi}, \\ w &\equiv z = r \cos \theta, \end{aligned}$$

which yields

$$g_M(\mathbf{r}) = \sum_{\alpha, \beta, \gamma=0}^M b_{\alpha\beta\gamma}^{(M)} u^\alpha v^\beta w^\gamma \quad (5.20)$$

$$= \sum_{l=0}^{3M} r^l \sum_{(\alpha, \beta, \gamma)} b_{\alpha\beta\gamma}^{(M)} e^{i(\alpha-\beta)\phi} \sin^{\alpha+\beta} \theta \cos^\gamma \theta. \quad (5.21)$$

In (5.21), the symbol  $\sum_{(\alpha, \beta, \gamma)}$  indicates taking the sums over combinations of  $\alpha, \beta, \gamma$  subject to the condition  $\alpha + \beta + \gamma = l$ . [Note that the sum over all  $l$  in effect removes the restriction on  $\alpha, \beta, \gamma$  and gives the same results as the original unrestricted sum in (5.20).]

We now restrict  $\mathbf{r}$  to the unit sphere by requiring that  $|\mathbf{r}| = 1$ , and introduce an index  $m = \alpha - \beta$ . The expression (5.21) is then rewritten in the form

$$g_M(\theta, \phi) = \sum_{l=0}^{3M} \sum_{(\alpha, \beta, \gamma)} b_{\alpha\beta\gamma}^{(M)} e^{im\phi} \sin^{\alpha+\beta-|m|} \theta \cos^\gamma \theta \sin^{|m|} \theta.$$

A trigonometric identity gives

$$\sin^{\alpha+\beta-|m|} \theta \cos^\gamma \theta = (1 - \cos^2 \theta)^{(\alpha+\beta-|m|)/2} \cos^\gamma \theta,$$

which is a polynomial in  $\cos \theta$  of maximum degree  $\alpha + \beta + \gamma - |m| = l - |m|$ , since  $\alpha + \beta - |m|$  is even (see the remark below). Denoting this polynomial by  $f_{lm}(\cos \theta)$ , we get

$$g_M(\theta, \phi) = \sum_{l=0}^{3M} \sum_m b_{lm}^{(M)} e^{im\phi} \sin^{|m|} \theta f_{lm}(\cos \theta). \quad (5.22)$$

*Remark.* That  $\alpha + \beta - |m|$  is even is seen by observing the identity

$$\alpha + \beta - |m| = m - |m| + 2\beta.$$

On the right-hand side,  $2\beta$  is even and

$$m - |m| = \begin{cases} 0 & \text{if } m \geq 0, \\ -2m & \text{if } m < 0. \end{cases}$$

The range of the summation over  $m$  still has to be specified. Recall that all the  $\alpha, \beta, \gamma$  are nonnegative integers subject to the condition that  $\alpha + \beta + \gamma = \ell \geq 0$ . This is illustrated schematically in Fig. 5.2, in which the point  $(\alpha, \beta, \gamma)$  must lie on the oblique face of the tetrahedron depicted in the  $\alpha\beta\gamma$  space. The line  $\alpha - \beta = m$  on the  $\gamma$  plane is shown as a solid line. In order for it to intersect the oblique face,  $m$  must satisfy the condition that

$$-\ell \leq m \leq \ell.$$

Therefore, the sum over  $m$  in (5.22) is restricted to  $|m| \leq l$ , and the last equation becomes

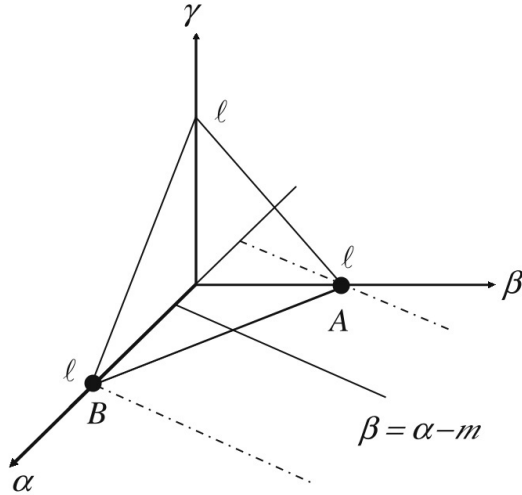
$$g_M(\theta, \phi) = \sum_{l=0}^{3M} \sum_{m=-l}^l b_{lm}^{(M)} Y_{lm}(\theta, \phi). \quad (5.23)$$

Here the sequence of functions

$$Y_{lm}(\theta, \phi) \equiv e^{im\phi} \sin^{|m|} \theta f_{lm}(\cos \theta), \quad (5.24)$$

where  $f_{lm}(\cos \theta)$  is a polynomial in  $\cos \theta$  of degree  $l - |m|$ , provides a uniform approximation to any continuous function defined on the unit sphere. The functions  $Y_{lm}$  are called **spherical harmonics**. Note that for a given  $l$ , there are  $2l + 1$  functions  $Y_{lm}$ .

The orthonormality of the set  $\{Y_{lm}\}$  is characterized by the relation



**Fig. 5.2.** The solid and dashed-dotted lines shown on the  $\gamma$ -plane indicate the relation  $\alpha - \beta = m$  for  $-\ell < m < \ell$  and  $m = \pm\ell$ , respectively. In order for the point  $(\alpha, \beta, \gamma)$  be on the oblique face of the tetrahedron, the condition  $-\ell \leq m \leq \ell$  should be satisfied so that the solid line intersects the line segment  $AB$  on the  $\gamma$ -plane

$$\int_0^{2\pi} d\phi \int_0^\pi \sin \theta d\theta Y_{l'm'}^*(\theta, \phi) Y_{lm}(\theta, \phi) = \delta_{ll'} \delta_{mm'},$$

which determines the functions  $Y_{lm}$  uniquely up to a phase factor.

General equations for the  $Y_{lm}$  are

$$Y_{lm}(\theta, \phi) = (-1)^m \left[ \frac{2l+1}{4\pi} \frac{(l-m)!}{(l+m)!} \right]^{1/2} P_l^m(\cos \theta) e^{im\phi}, \quad m \geq 0$$

$$Y_{l,-m}(\theta, \phi) = (-1)^m Y_{lm}^*(\theta, \phi), \quad m \geq 0,$$

where

$$\begin{aligned} P_l^m(x) &= (1-x^2)^{m/2} \frac{d^m}{dx^m} P_l(x) \\ &= \frac{1}{2^\ell \ell!} (1-x^2)^{m/2} \frac{d^{\ell+m}}{dx^{\ell+m}} (x^2-1)^\ell, \quad m \geq 0, \end{aligned} \quad (5.25)$$

are called the **associated Legendre functions**.

*Remark.*

1. The normalization constant of the  $Y_{lm}$  follows immediately from the orthonormality relations for the associated Legendre functions:

$$\int_{-1}^1 P_l^m(x) P_{l'}^m(x) dx = \frac{(l+m)!}{(l-m)!} \frac{2}{2l+1} \delta_{ll'}.$$

There is, of course, a free choice of phase factor; ours is a common choice in the physics literature. However, one must be careful because different authors choose different phase factors for the spherical harmonics.

2. We should note that the associated Legendre functions  $P_l^m(x)$  are not another orthonormal set of polynomials on  $[-1, 1]$ . In fact, they are not polynomials at all as is clearly seen in equation (5.25).

## Exercises

1. Find the normalized Legendre polynomials  $\tilde{P}_n(x)$ .

**Solution:** Using equation (5.13), we write the normalized Legendre polynomials  $\tilde{P}_n(x)$  as

$$\tilde{P}_n(x) = \sqrt{\frac{2n+1}{2n}} P_n(x), \quad n = 0, 1, 2, \dots \quad \clubsuit$$

2. Derive the explicit form of each function:  $Y_{00}$ ,  $Y_{11}$ ,  $Y_{10}$ , and  $Y_{1,-1}$ .

**Solution:** It follows from (5.24) that for  $l = m = 0$ , we obtain  $Y_{00} = \sqrt{1/4\pi}$ . If  $l = 1$ , then  $m$  can equal  $-1$ ,  $0$ , or  $+1$ . Recalling

that  $f_{lm}(\cos \theta)$  is a polynomial in  $\cos \theta$  of degree  $l - |m|$ , we obtain  $Y_{10} = c_1 \cos \theta + c_2$ ,  $Y_{11} = c_3 e^{i\phi} \sin \theta$ ,  $Y_{1,-1} = c_4 e^{-i\phi} \sin \theta$ . The constants  $c_1, c_2, c_3, c_4$  are determined by imposing orthonormality. For instance,

$$\begin{aligned} \int_0^{2\pi} d\phi \int_0^\pi \sin \theta Y_{00}^* Y_{10} d\theta &= \sqrt{\pi} \int_0^\pi d\theta \sin \theta (c_1 \cos \theta + c_2) \\ &= \delta_{01} \delta_{00} = 0, \\ \int_0^{2\pi} d\phi \int_0^\pi \sin \theta |Y_{10}|^2 d\theta &= 2\pi \int_0^\pi d\theta [\sin \theta (c_1 \cos \theta + c_2)]^2 \\ &= \delta_{10} \delta_{10} = 1, \end{aligned}$$

which result in  $c_1 = \sqrt{3/(4\pi)}$  and  $c_2 = 0$ . Similarly, it follows that  $c_3 = -c_4 = -\sqrt{3/8\pi}$ . We choose the minus sign with the convention to be adopted later. Therefore, the first few members of the set  $\{Y_{lm}\}$  are

$$\begin{aligned} Y_{00} &= \sqrt{\frac{1}{4\pi}}, & Y_{11} &= -\sqrt{\frac{3}{8\pi}} e^{i\phi} \sin \theta, \\ Y_{10} &= \sqrt{\frac{3}{4\pi}} \cos \theta, & Y_{1,-1} &= \sqrt{\frac{3}{8\pi}} e^{-i\phi} \sin \theta. \quad \clubsuit \end{aligned}$$

**3.** From the generating function of Legendre polynomials determine that

- (i)  $P_n(1) = 1$ ,  $P_n(-1) = (-1)^n$ ,
- (ii)  $P_{2n}(0) = (-1)^n \frac{(2n-1)!!}{(2n)!!}$ ,  $P_{2n+1}(0) = 0$  with  $(-1)!! = 1$ ,
- (iii)  $\int_0^1 x^n P_n(x) dx = \frac{2^n (n!)^2}{(2n+1)!}$ .

**Solution:** We use the equation

$$(1 - 2tx + t^2)^{-1/2} = \sum_{n=0}^{\infty} P_n(x) t^n.$$

(i) For  $x = 1$ , we have

$$\frac{1}{1-t} = \sum_{n=0}^{\infty} t^n = \sum_{n=0}^{\infty} P_n(1) t^n,$$

which yields  $P_n(1) = 1$ . Similarly for  $x = -1$ , we obtain

$$\frac{1}{1+t} = \sum_{n=0}^{\infty} (-1)^n t^n = \sum_{n=0}^{\infty} P_n(-1) t^n,$$

which gives  $P_n(-1) = (-1)^n$ .

(ii) For  $x = 0$ , we have

$$\left( \frac{1}{1+t^2} \right)^{1/2} = \sum_{n=0}^{\infty} (-1)^n \frac{(2n-1)!!}{(2^n n!)} t^{2n} = \sum_{n=0}^{\infty} P_n(0) t^n.$$

Then, we have the desired result.

(iii) We get the relation by performing an integration in parts  $n$  times. ♣

4. Show that the Coulomb potential at  $\mathbf{r} = \mathbf{r}_0$  experienced from the unit charge at  $z = a$  on the  $z$ -axis is given by

$$V(\mathbf{r}_0) = \frac{1}{4\pi a} \sum_{n=0}^{\infty} \left( \frac{r_0}{a} \right)^n P_n(\cos \theta),$$

where  $\theta$  is the angle between the  $z$  axis and the vector  $\mathbf{r}_0$  and  $a$  satisfies the condition  $r_0 < a$ .

**Solution:** Using the generating function of Legendre polynomials, we have

$$\begin{aligned} V(\mathbf{r}_0) &= \frac{1}{4\pi\epsilon_0} \frac{1}{|\mathbf{r}_0 - \mathbf{a}|} = \frac{1}{4\pi\epsilon_0} \frac{1}{\sqrt{r_0^2 + a^2 - 2ar_0 \cos \theta}} \\ &= \frac{1}{4\pi\epsilon_0} \sum_{n=0}^{\infty} \left( \frac{r_0}{a} \right)^n P_n(\cos \theta). \end{aligned}$$

This series converges because  $r_0 < a$  and  $|P_n(\cos \theta)| \leq 1$ . ♣

## 5.2 Classification of Orthonormal Functions

### 5.2.1 General Rodrigues Formula

In the previous section we saw that several kinds of orthonormal polynomials can be produced through the Gram-Schmidt orthogonalization process by starting with  $1, x, x^2, \dots$ . However, there is a more elegant approach that applies to most polynomials of interest to physicists. This section describes this approach, which is based on the **Rodrigues formula** and classifies various orthogonal polynomials in terms of the parameters involved in the formula.

♠ **General Rodrigues formula:**

$$Q_n(x) = \frac{1}{K_n w(x)} \frac{d^n}{dx^n} [w(x)s^n(x)] \quad (n = 0, 1, 2, \dots), \quad (5.26)$$

where it is assumed that

1.  $Q_1(x)$  is a first-degree polynomial in  $x$ .
2.  $s(x)$  is a polynomial in  $x$  of degree no more than 2 with real roots.
3.  $w(x)$  is real, positive, and integrable in the interval  $[a, b]$  and satisfies the boundary condition

$$w(a)s(a) = w(b)s(b) = 0.$$

Equation (5.26) under the three conditions noted above provides the sequence of functions  $(Q_0(x), Q_1(x), Q_2(x), \dots)$  that forms an orthogonal set of polynomials on the interval  $[a, b]$  with a weight function  $w(x)$ , which can be normalized by a suitable choice of constants  $K_n$ . For historical reasons, different polynomial functions are normalized differently, which is why  $K_n$  is introduced here. In the meantime, we omit denoting  $K_n$  without loss of generality.

♠ **Theorem:**

The function  $Q_n(x)$  defined by (5.26) is a polynomial in  $x$  of the  $n$ th degree and satisfies the orthogonality relation on the interval  $[a, b]$  with weight  $w(x)$ :

$$\int_a^b p_m(x) Q_n(x) w(x) dx = 0 \quad (m < n), \quad (5.27)$$

where  $p_m(x)$  is an arbitrary polynomial of degree  $m < n$ .

*Proof* From hypothesis, we have

$$\left. \frac{d^m}{dx^m} [w(x)s^n(x)] \right|_{x=a \text{ or } b} = 0 \quad (\text{if } m < n) \quad (5.28)$$

and

$$\frac{d^m}{dx^m} [w(x)s^n(x)p_{(\leq k)}(x)] = w(x)s^{n-m}(x)p_{(\leq k+m)}(x), \quad (5.29)$$

where the symbol  $p_{(\leq k)}(x)$  denotes an arbitrary polynomial in  $x$  of degree  $\leq k$ . Then, integrating (5.27) by parts  $n$  times, we obtain for  $m < n$ ,

$$\begin{aligned}
\int_a^b p_m(x) Q_n(x) w(x) dx &= \int_a^b p_m(x) \frac{d^n}{dx^n} [w(x) s^n(x)] dx \\
&= \int_a^b w(x) s^n(x) \frac{d^n}{dx^n} p_m(x) dx = 0, \quad (5.30)
\end{aligned}$$

where we used (5.26) and (5.28). Next we examine whether or not  $Q_n$  is a polynomial of degree  $n$ . Set  $n = m$  and  $k = 0$  in (5.29) to obtain

$$\frac{1}{w(x)} \frac{d^n}{dx^n} [w(x) s^n(x)] = Q_n(x) = p_{(\leq n)}(x),$$

which indicates that  $Q_n(x)$  is a polynomial of degree no more than  $n$ . We thus tentatively write

$$Q_n(x) = p_{(\leq n-1)}(x) + a_n x^n, \quad (5.31)$$

and would like to show that  $a_n \neq 0$ . Multiplying both parts of (5.31) by  $Q_n(x)w(x)$  followed by integrating on  $[a, b]$  yields

$$\begin{aligned}
\int_a^b [Q_n(x)]^2 w(x) dx &= \int_a^b p_{(\leq n-1)}(x) Q_n(x) w(x) dx + a_n \int_a^b x^n Q_n(x) w(x) dx \\
&= a_n \int_a^b x^n Q_n(x) w(x) dx,
\end{aligned}$$

where we used (5.30). This clearly proves that  $a_n \neq 0$ , i.e., that  $Q_n(x)$  is a polynomial of the  $n$ th degree. ♣

### 5.2.2 Classification of the Polynomials

In what follows, we classify the orthogonal polynomials that are derived from the Rodrigues formula (5.26) the three conditions according to noted earlier. By the condition **1** associated with (5.26),  $Q_1(x)$  is a first-degree polynomial, and we can define it as

$$Q_1(x) = -\frac{x}{K_1}. \quad (5.32)$$

Then the Rodrigues formula (5.26) reads

$$\frac{1}{w} \frac{dw}{dx} = -\frac{x + (ds/dx)}{s}. \quad (5.33)$$

Recall that  $s(x)$  can be the zeroth-, first-, or second-degree polynomial. In each case, we can find an appropriate weight function  $w(x)$  that satisfies the differential equation (5.33) as well as the boundary condition **3**:

$$w(a)s(a) = w(b)s(b) = 0. \quad (5.34)$$

Such discussions determine the explicit forms of possible functions  $s(x)$  and  $w(x)$  under conditions **1–3** in Sect. 5.2.1 and then allow classification of all of the orthogonal polynomials provided by the general Rodrigues formula described below.

### Hermite polynomials:

We first consider the case that  $s(x)$  is a zeroth-degree polynomial, i.e., a constant given by

$$s(x) = \alpha.$$

Equation (5.33) takes the form

$$\frac{1}{w} \frac{dw}{dx} = -\frac{x}{\alpha}$$

and has the solution

$$w(x) = A \exp\left(-\frac{x^2}{2\alpha}\right) \quad \text{with a constant } A. \quad (5.35)$$

The product  $w(x)s(x)$  vanishes only at  $x = \pm\infty$ , provided that  $\alpha > 0$ . To satisfy the conditions in (5.34), we have to set

$$a = -\infty, \quad b = +\infty.$$

The constants  $A$  and  $\alpha$  affect only the multiplicative factor in front of each polynomial. Thus, without loss of generality, we can take  $\alpha = 1$  and  $A = 1$ , which yields

$$w = e^{-x^2}.$$

The complete orthonormal polynomials corresponding to this case are known as Hermite polynomials, designated by  $H_n(x)$ , and satisfy the orthonormal condition

$$\int_{-\infty}^{\infty} e^{-x^2} H_m(x) H_n(x) dx = \delta_{mn}.$$

### Laguerre polynomials:

Next we let  $s(x)$  be a polynomial of the first degree, such as

$$s(x) = \beta(x - \alpha).$$

The Rodrigues formula (5.26) now becomes

$$\frac{1}{w} \frac{dw}{dx} = -\frac{x + \beta}{\beta(x - \alpha)},$$

which has the solution

$$w(x) = \text{const.} \times (x - \alpha)^\nu e^{-x/\beta},$$

where

$$\nu \equiv -\frac{\alpha + \beta}{\beta}.$$

If  $\beta > 0$  and  $\nu > -1$ , then  $s(x)w(x)$  vanishes at  $x = \alpha$  and  $x = +\infty$ , and  $w(x)$  is integrable in the interval  $[\alpha, +\infty)$ . The simplest choice is therefore to take  $\alpha = 0$  and  $\beta = 1$ , which yields

$$w = x^\nu e^{-x}, \quad a = 0, \quad b = +\infty.$$

These choices result in the Laguerre polynomials, commonly denoted by  $L_n^\nu(x)$ , whose orthonormality relation is given by

$$\int_0^\infty x^{-\nu} L_m^\nu(x) L_n^\nu(x) dx = \delta_{mn} \quad \text{with } \nu > -1.$$

### Jacobi polynomials:

Finally, let us take

$$s(x) = \gamma(x - \alpha)(\beta - x), \quad \beta > \alpha.$$

Here we assume that  $s(x)$  has two distinct roots. [If  $s(x)$  has a double root, the boundary condition (5.34) cannot be satisfied, since in this case the function  $s(x)w(x)$  cannot vanish at more than one point.] The Rodrigues formula (5.26) now reads

$$\frac{1}{w} \frac{dw}{dx} = -\frac{x + \gamma(\beta - x) - \gamma(x - \alpha)}{\gamma(x - \alpha)(\beta - x)},$$

which has the solution

$$w(x) = \text{const.} \times (x - \alpha)^\mu (\beta - x)^\nu,$$

with

$$\mu \equiv -\frac{\alpha + \beta}{\beta} \quad \text{and} \quad \nu \equiv \frac{1 - \gamma}{\gamma} - \frac{\alpha}{\gamma(\beta - \alpha)}.$$

If  $\mu > -1$  and  $\nu > -1$ , then  $s(x)w(x)$  vanishes at  $x = \alpha$  and  $x = \beta$ , and  $w(x)$  is integrable on the interval  $[\alpha, \beta]$ . With the replacement

$$\frac{2x - \alpha - \beta}{\beta - \alpha} \rightarrow x,$$

apart from multiplicative factors, we obtain

$$w = (1-x)^\nu(1+x)^\mu \text{ with } \nu, \mu > -1, \quad a = -1, \quad b = +1.$$

The corresponding complete orthonormal polynomials are called the Jacobi polynomials  $G_n^{\mu,\nu}(x)$ , and satisfy the relation

$$\int_{-1}^1 (1-x)^{-\nu}(1+x)^{-\mu} G_m^{\mu,\nu}(x) G_n^{\mu,\nu}(x) dx = \delta_{mn} \quad \text{with } \nu, \mu > -1.$$

**Remark.** Jacobi polynomials can be divided into subcategories depending on the values of  $\mu$  and  $\nu$ . The most common and widely used in mathematical physics are collected in Table 5.1.

**Table 5.1.** Special cases of Jacobi polynomials

$\mu$	$\nu$	$w(x)$	Polynomial
$\lambda - 1/2$	$\lambda - 1/2$	$(1-x^2)^{\lambda-1/2}$	Gegenbauer, $C_n^\lambda(x)$ . $\lambda > -1/2$
0	0	1	Legendre, $P_n(x)$
$-1/2$	$-1/2$	$(1-x^2)^{-1/2}$	Chebyshev of the first kind, $T_n(x)$
$1/2$	$1/2$	$(1-x^2)^{1/2}$	Chebyshev of the second kind, $U_n(x)$

### 5.2.3 The Recurrence Formula

We now show that all the orthogonal polynomials derived from the Rodrigues formula (5.26) satisfy the following relation:

**♠ Recurrence formula:**

$$Q_{n+1}(x) = (a_n x + b_n) Q_n(x) - c_n Q_{n-1}(x), \quad (n = 1, 2, \dots) \quad (5.36)$$

where the constants  $a_n$ ,  $b_n$ , and  $c_n$  depend on the class of polynomials considered.

*Proof* The only property needed for the proof of (5.36) is the orthogonality relation:

$$\int_a^b Q_n(x) p_{(<n)}(x) w(x) dx = 0, \quad (5.37)$$

where the symbol  $p_{(<n)}(x)$  denotes an arbitrary polynomial in  $x$  of degree less than  $n$ . For convenience, we introduce the following notation:

$$\begin{aligned} \xi_n &= \text{coefficient of } x^n \text{ in } Q_n(x), \\ \eta_n &= \text{coefficient of } x^{n-1} \text{ in } Q_n(x), \end{aligned} \quad (5.38)$$

$$I_n = \int_a^b Q_n^2(x) w(x) dx. \quad (5.39)$$

It then follows that

$$Q_{n+1}(x) - \frac{\xi_{n+1}}{\xi_n} x Q_n(x) = \sum_{i=0}^n r_i^{(n)} Q_i(x)$$

because the left-hand side is a polynomial of degree  $\leq n$ ;  $r_i^{(n)}$  are appropriate constants determined by the left-hand side. Multiplying both sides by  $wQ_m$ , taking  $m$  equal to  $0, 1, 2, \dots, n-2$  successively, and using the orthogonality relation (5.37), we obtain

$$r_m^{(n)} = 0 \quad \text{for } m = 0, 1, 2, \dots, n-2.$$

Thus

$$Q_{n+1}(x) - \frac{\xi_{n+1}}{\xi_n} x Q_n(x) = r_n^{(n)} Q_n(x) + r_{n-1}^{(n)} Q_{n-1}(x), \quad (5.40)$$

which is the recurrence formula we are looking for. ♣

#### 5.2.4 Coefficients of the Recurrence Formula

We now have to find the constants  $r_n^{(n)}$  and  $r_{n-1}^{(n)}$  in (5.40). In view of the orthogonality relation (5.37), we have

$$I_n = \int_a^b Q_n^2(x) w(x) dx = \xi_n \int_a^b Q_n(x) x^n w(x) dx. \quad (5.41)$$

Multiplying (5.40) by  $wQ_{n-1}$  and integrating, we obtain

$$\begin{aligned} I_{n-1} r_{n-1}^{(n)} &= -\frac{\xi_{n+1}}{\xi_n} \int_a^b Q_n(x) Q_{n-1}(x) x w(x) dx \\ &= -\frac{\xi_{n+1}}{\xi_n} \cdot \frac{\xi_{n-1}}{\xi_n} \int_a^b Q_n(x) \xi_n x^n w(x) dx \\ &= -\frac{\xi_{n+1} \xi_{n-1}}{\xi_n^2} I_n. \end{aligned}$$

Therefore,

$$r_{n-1}^{(n)} = -\frac{I_n}{I_{n-1}} \cdot \frac{\xi_{n+1} \xi_{n-1}}{\xi_n^2}. \quad (5.42)$$

Substituting this into (5.40) and comparing the coefficients of  $x^n$  on both sides, yields

$$r_n^{(n)} = -\frac{\eta_{n+1}}{\xi_n} \cdot \frac{\xi_{n+1} \eta_n}{\xi_n^2}. \quad (5.43)$$

Finally, it follows from (5.40)–(5.43) that the coefficients  $a_n$ ,  $b_n$ , and  $c_n$  defined in (5.40) become

$$\begin{aligned}
a_n &= \frac{\xi_{n+1}}{\xi_n}, \\
b_n &= \frac{\xi_{n+1}}{\xi_n} \left( \frac{\eta_{n+1}}{\xi_{n+1}} - \frac{\eta_n}{\xi_n} \right), \\
c_n &= \frac{I_n}{I_{n-1}} \cdot \frac{\xi_{n+1}\xi_{n-1}}{\xi_n^2}.
\end{aligned} \tag{5.44}$$

The constants  $\xi_n$  and  $\eta_n$  can, in principle, be found from the Rodrigues formula once the functions  $s(x)$  and  $w(x)$  as well as the constants  $K_n$  have been fixed. The constants  $I_n$ , which determine the normalization of the polynomials, are given by

$$I_n = \frac{(-1)^n \xi_n n!}{K_n} \int_a^b s(x)^n w(x) dx.$$

This follows immediately from the Rodrigues formula if we integrate  $n$  times by parts the integral

$$\begin{aligned}
I_n &= \int_a^b Q_n(x)^2 w(x) dx = \xi_n \int_a^b Q_n(x) x^n w(x) dx \\
&= \frac{\xi_n}{K_n} \int_a^b x^n \frac{d^n}{dx^n} [s(x)^n w(x)] dx.
\end{aligned}$$

Although the explicit form of the coefficients given in (5.44) seems rather complicated, the corresponding recurrence relation for a specific orthogonal polynomial simplifies it considerably.

### 5.2.5 Roots of Orthogonal Polynomials

Consider the recurrence formula (5.36) in which the polynomials  $Q_n(x)$  are normalized, and from (5.39)  $I_n = 1$  ( $n = 0, 1, 2, \dots$ ). After some rearrangement, the equation takes the form

$$xQ_{n-1}(x) = \frac{\xi_{n-1}}{\xi_n} Q_n(x) + \frac{\xi_{n-2}}{\xi_{n-1}} Q_{n-2}(x) + \beta_{n-1} Q_{n-1}(x),$$

where

$$\beta_{n-1} = \frac{\eta_{n-1}}{\xi_{n-1}} - \frac{\eta_n}{\xi_n}.$$

The matrix form is given by

$$\begin{aligned}
x \begin{pmatrix} Q_0 \\ Q_1 \\ Q_2 \\ \dots \\ \dots \\ Q_{N-1} \end{pmatrix} &= \begin{pmatrix} \beta_0 & \xi_0/\xi_1 & 0 & \dots & \dots & 0 \\ \xi_0/\xi_1 & \beta_1 & \xi_1/\xi_2 & \dots & \dots & 0 \\ 0 & \xi_1/\xi_2 & \beta_2 & \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \dots & \beta_{N-1} \end{pmatrix} \begin{pmatrix} Q_0 \\ Q_1 \\ Q_2 \\ \dots \\ \dots \\ Q_{N-1} \end{pmatrix} \\
&+ \begin{pmatrix} 0 \\ 0 \\ 0 \\ \dots \\ \dots \\ (\xi_{N-1}/\xi_N)Q_N \end{pmatrix},
\end{aligned}$$

which gives the eigenvalue equations provided that  $\{x_i\}$  are the roots of the polynomial equation  $Q_N(x) = 0$  such that

$$J\mathbf{R}(x_i) = x_i\mathbf{R}(x_i),$$

where the column vector  $\mathbf{R}(x_i)$  is defined by

$$\mathbf{R}(x_i) = [Q_0(x_i), Q_1(x_i), Q_{N-1}(x_i)].$$

Thus, the eigenvalues of the  $N \times N$  matrix  $J$  are the zeros of  $Q_N(x)$ . The matrix is called the **Jacobi matrix** associated with the sequence  $\{Q_n(x)\}$ . Since  $J$  is symmetric, the eigenvalues  $\{x_i\}$  are real. We thus have proved the following theorem:

♠ **Theorem:**

The eigenvalues  $\{x_i\}$  ( $i = 1, 2, \dots, N$ ) of the matrix  $J$  are the zeros of  $Q_N(x)$ . The eigenvector belonging to  $x_i$  is  $\mathbf{R}(x_i) = [Q_0(x_i), Q_1(x_i), Q_{N-1}(x_i)]$ .

### 5.2.6 Differential Equations Satisfied by the Polynomials

Historically, most orthogonal polynomials were discovered as solutions of differential equations. Here we give a single generic differential equation that is satisfied by all the polynomials  $Q_n$ .

♠ **Theorem:**

All of the orthogonal polynomials  $Q_n(x)$  derived from the general Rodrigues formula (5.26) satisfy the differential equation

$$\frac{d}{dx} \left( sw \frac{dQ_n}{dx} \right) = -\lambda_n w Q_n,$$

with the constant

$$\lambda_n = -n \left( K_1 \frac{dQ_1}{dx} + \frac{n-1}{2} \frac{d^2 s}{dx^2} \right).$$

*Proof* Since  $dQ_n(x)/dx$  is a polynomial of degree  $\leq (n-1)$ , it follows from (5.29) that the function

$$\frac{1}{w} \frac{d}{dx} \left[ s(x)w(x) \frac{dQ_n}{dx} \right]$$

is a polynomial of degree  $\leq n$ . Thus, we can write

$$\frac{1}{w} \frac{d}{dx} \left[ s(x)w(x) \frac{dQ_n}{dx} \right] = - \sum_{i=1}^n \lambda_n^{(i)} Q_i(x), \quad (5.45)$$

where the  $\lambda_n^{(i)}$  are undetermined constants. Multiplying both sides of (5.45) by  $wQ_m$  and integrating, we get

$$\int_a^b Q_m(x) \frac{d}{dx} \left[ s(x)w(x) \frac{dQ_n}{dx} \right] dx = -\lambda_n^{(m)} I_m. \quad (5.46)$$

Here  $I_m$  is an integral given by (5.39). Integrating by parts, for  $m < n$  the left-hand side of (5.46) yields

$$\begin{aligned} & \int_a^b Q_m(x) \frac{d}{dx} \left[ s(x)w(x) \frac{dQ_n}{dx} \right] dx \\ &= - \int_a^b s(x)w(x) \frac{dQ_n}{dx} \frac{dQ_m}{dx} dx \\ &= \int_a^b w(x)Q_n(x) \left[ \frac{1}{w} \frac{d}{dx} \left( s(x)w(x) \frac{dQ_m}{dx} \right) \right] dx \\ &= 0. \end{aligned}$$

We have used the condition that  $s(a)w(a) = s(b)w(b) = 0$ , which is assumption **3** in Sect. 5.2.1. We also used the fact that  $Q_n(x)$  is orthogonal to any polynomial of degree  $< n$ . Consequently, we arrive at the result

$$\lambda_n^{(m)} = 0, \quad \text{for } m < n.$$

Setting

$$\lambda_n^{(n)} = \lambda_n,$$

for simplicity, we can rewrite (5.45) in the form

$$\frac{d}{dx} \left[ s(x)w(x) \frac{dQ_n}{dx} \right] = -w(x)\lambda_n Q_n(x), \quad (5.47)$$

which is the differential equation satisfied by a polynomial  $Q_n(x)$ . The constant  $\lambda_n$  can be found by setting  $m = n$  in (5.46) and integrating, as we demonstrate later in Exercise 4.

### 5.2.7 Generating Functions (I)

As a matter of fact, all the orthogonal polynomials  $Q_n(x)$  discussed thus far can be generated from a single function  $g(t, x)$  of two variables by repeated differentiation with respect to  $t$ . Called a **generating function**, it plays a significant role in many areas of mathematics. Here we study the essence of generating functions together with several examples by which we can derive specific orthogonal polynomials.

A formal definition of generating functions is given below.

#### ♠ Generating function:

Assume a (finite or infinite) convergent power series

$$\gamma(t) \equiv \sum_k f_k t^k.$$

The  $\gamma(t)$  is called a **generating function** for the sequence of coefficients  $f_1, f_2, \dots, f_n, \dots$ .

Clearly, all the coefficients  $f_n$  are obtained from differentiating  $\gamma(t)$  as given by

$$f_n = \frac{1}{n!} \frac{d^n \gamma(t)}{dt^n}.$$

For orthogonal polynomials, generating functions are assumed to take the form

$$g(t, x) = \sum_{n=0}^{\infty} A_n Q_n(x) t^n, \quad (5.48)$$

where  $Q_n(x)$  is an orthogonal polynomial associated with  $g(t, x)$ , and the  $A_n$  are appropriate constants. The explicit form of  $g(t, x)$  can often be derived using the Rodrigues formula and **Cauchy's integral formula** (see Sect. 7.3.1).

Remember that the latter formula determines an  $n$ th-order derivative of a function  $f(z)$  as

$$\frac{d^n}{dz^n} f(z) = \frac{n!}{2\pi i} \int_C \frac{f(\zeta) d\zeta}{(\zeta - z)^{n+1}},$$

where  $f(z)$  is **analytic** within the closed contour  $C$ . (See Sect. 7.1.2 for a definition of **analytic functions**.) Applying this to the Rodrigues formula for, say, Hermite polynomials  $H_n(x)$ , we obtain

$$H_n(x) = (-1)^n e^{x^2/2} \frac{d^n}{dx^n} e^{-x^2/2} = (-1)^n e^{x^2/2} \frac{n!}{2\pi i} \oint_C \frac{e^{-\zeta^2/2} d\zeta}{(\zeta - x)^{n+1}}.$$

We then try to sum the series as

$$\sum_{n=0}^{\infty} \frac{H_n(x)}{n!} t^n = \frac{e^{x^2/2}}{2\pi i} \oint_C e^{-\zeta^2/2} \left[ \sum_{n=0}^{\infty} \frac{(-1)^n t^n}{(\zeta - x)^{n+1}} \right] d\zeta = \frac{e^{x^2/2}}{2\pi i} \oint_C \frac{e^{-\zeta^2/2} d\zeta}{\zeta - x + t},$$

where we require that the point  $x - t$  be inside the contour. Finally we evaluate the above integral and find

$$e^{tx - (t^2/2)} = \sum_{n=0}^{\infty} \frac{H_n(x) t^n}{n!}.$$

Comparing this last equation with (5.48), we see that  $e^{tx - (t^2/2)}$  is the generating function associated with Hermite polynomials  $H_n(x)$ . Similarly, we can derive the generating function for Laguerre polynomials as

$$\frac{e^{-tx/(1-t)}}{(1-t)^{1+\alpha}} = \sum_{n=0}^{\infty} L_n^\alpha(x) t^n.$$

### 5.2.8 Generating Functions (II)

There is an alternative way to determine a generating function, which is based on the recurrence formula for a particular polynomial. To see this, we try to find the generating function of the Legendre polynomials that satisfies the following recursion formula:

$$(n+1)P_{n+1}(x) - (2n+1)xP_n(x) + nP_{n-1}(x) = 0,$$

with  $P_0(x) = 1$ ,  $P_1(x) = x$ , and for convenience we set  $P_{-1}(x) = 0$ . We seek an expression in closed form for

$$g(t, x) = \sum_{n=0}^{\infty} P_n(x) t^n.$$

First we note that

$$\begin{aligned}\frac{\partial g}{\partial t} &= \sum_{n=0}^{\infty} n P_n(x) t^{n-1} = \sum_{n=0}^{\infty} (n+1) P_{n+1}(x) t^n \\ &= \sum_{n=0}^{\infty} [(2n+1)x P_n(x) - n P_{n-1}(x)] t^n.\end{aligned}$$

By straightforward rearrangement we find that

$$\frac{\partial g}{\partial t} = xg(t, x) + 2tx \frac{\partial g}{\partial t} - tg(t, x) - t^2 \frac{\partial g}{\partial t},$$

which leads to the partial differential equation

$$\frac{1}{g} \frac{\partial g}{\partial t} = \frac{x-t}{1-2tx+t^2}.$$

Coupled with the initial condition  $g(0, x) = 1$  we finally have

$$\frac{1}{\sqrt{1-2tx+t^2}} = \sum_{n=0}^{\infty} P_n(x) t^n.$$

Generating functions for other orthogonal polynomials are given in Appendix D.

## Exercises

1. Find the recurrence formula for normalized polynomials  $\tilde{Q}_n(x)$ .

**Solution:** When the polynomials are normalized, we have  $I_n = 1$  ( $n = 0, 1, 2, \dots$ ) from (5.39). The recurrence formula (5.36) is

$$\tilde{Q}_{n+1}(x) = (a_n x + b_n) \tilde{Q}_n(x) - \frac{a_n}{a_{n-1}} \tilde{Q}_{n-1}(x). \quad \clubsuit$$

2. Assume that a sequence of orthogonal polynomials satisfies

$$Q_{n+1}(x) = [(n+1)x + 1] Q_n(x) - 3(n+1)Q_{n-1}(x).$$

Find the normalized constants for  $Q_n(x)$  defined by  $\tilde{Q}_n(x) = \lambda Q_n(x)$ , where  $\tilde{Q}_n(x)$  are normalized polynomials.

**Solution:** We denote the normalized polynomials as  $\tilde{Q}_n(x) = \lambda Q_n(x)$ , where the constants  $\lambda_n$  ( $n = 0, 1, 2, \dots$ ) are to be found. Substituting  $\tilde{Q}_n(x)$  into the given formula, we have

$$\tilde{Q}_{n+1}(x) = \frac{\lambda_{n+1}}{\lambda_n} [(n+1)x + 1] \tilde{Q}_n(x) - 3(n+1) \frac{\lambda_{n+1}}{\lambda_n} \tilde{Q}_{n-1}(x).$$

Comparing this with the normalized recurrence formula from Exercise 1, we have the relation  $(3\lambda_n)/\lambda_{n-1} = \lambda_{n-1}/(n\lambda_n)$ , which yields  $\lambda_n = \lambda_{n-1}/\sqrt{3n}$ . This relation gives the normalization constants of the form

$$\lambda_n = \frac{1}{3^{n/2}} (n!)^{1/2} \lambda_0. \quad \clubsuit$$

3. Find the recurrence formula for Hermite and Legendre polynomials.

**Solution:** For Hermite and Legendre polynomials, (5.36) reads

$$H_{n+1}(x) = 2xH_n(x) - 2nH_{n-1}(x) \quad (5.49)$$

and

$$(n+1)P_{n+1}(x) = (2n+1)xP_n(x) - nP_{n-1}(x), \quad (5.50)$$

respectively. See Appendix D for the recurrence relations associated with the other polynomials we have discussed.  $\clubsuit$

4. Determine the constants  $\lambda_n$  given in (5.47).

**Solution:** Setting  $m = n$  on the left-hand side of (5.46), we obtain

$$\int_a^b Q_n(x) \frac{d}{dx} \left[ s(x)w(x) \frac{dQ_n}{dx} \right] dx \quad (5.51)$$

$$\begin{aligned} &= \int_a^b Q_n(x) \left[ \frac{d(sw)}{dx} \frac{dQ_n}{dx} + s(x)w(x) \frac{d^2Q_n}{dx^2} \right] dx \\ &= \int_a^b w(x)Q_n(x) \left[ K_1Q_1(x) \frac{dQ_n}{dx} + s(x) \frac{d^2Q_n}{dx^2} \right] dx. \end{aligned} \quad (5.52)$$

Here we used the relation  $d(sw)/dx = wK_1Q_1$  [set  $n = 1$  in the general Rodrigues formula (5.26).] The orthogonality of  $Q_n(x)$  means that only the  $n$ th power of  $x$  in the square brackets contributes to the integral in the last line of (5.52). [See (5.30) for details of the orthogonal property of  $Q_n(x)$ .] We then set up the following expressions:

$$\begin{aligned}
s(x) &= ax^2 + bx + c, \\
Q_n(x) &= \xi_n x^n + \xi_{n-1} x^{n-1} + \cdots, \\
Q_1(x) &= \eta_1 + \eta_0,
\end{aligned}$$

which result in

$$K_1 Q_1 \frac{dQ_n}{dx} = K_1 \eta_1 n \xi_n x^n + (\text{const.}) \times x^{n-1} + \cdots$$

and

$$s \frac{d^2 Q_n}{dx^2} = an(n-1) \xi_n x^n + (\text{const.}) \times x^{n-1} + \cdots.$$

Thus the relevant terms in the square brackets in the last line in (5.52) become

$$\left[ n K_1 \frac{dQ_1}{dx} + \frac{1}{2} \frac{ds^2}{dx^2} n(n-1) \right] \xi_n x^n,$$

where we used  $\eta_1 = dQ_1/dx$  and  $a = (1/2)(ds^2/dx^2)$ , and we get

$$\begin{aligned}
& \int_a^b Q_n(x) \frac{d}{dx} \left[ s(x) w(x) \frac{dQ_n}{dx} \right] dx \\
&= \left[ n K_1 \frac{dQ_1}{dx} + \frac{n}{2} (n-1) \frac{d^2 s}{dx^2} \right] \int_a^b w(x) Q_n(x) (\xi_n x^n) dx \\
&= n \left( K_1 \frac{dQ_1}{dx} + \frac{n-1}{2} \frac{d^2 s}{dx^2} \right) I_n.
\end{aligned}$$

Comparing this with (5.46), gives us

$$\lambda_n = -n \left( K_1 \frac{dQ_1}{dx} + \frac{n-1}{2} \frac{d^2 s}{dx^2} \right). \quad \clubsuit$$

## 5.3 Chebyshev Polynomials

### 5.3.1 Minimax Property

Thus far we have seen that every real function  $f(x)$  defined in a certain interval (finite or infinite) can be approximated in the mean by appropriate orthogonal polynomial  $\{Q_n(x)\}$  as

$$f(x) \simeq \sum_{i=0}^n c_i Q_i(x). \quad (5.53)$$

The coefficients  $c_i$  are determined formally by using the orthogonality of the polynomials in question. The striking advantage of such polynomial approximations is that an improvement in the approximation through addition of an extra term  $c_{n+1} Q_{n+1}(x)$  does not affect the previously obtained coefficients,  $c_0, c_1, \dots, c_n$ .

In principle, any polynomial that we discussed in Sect. 5.2 can be approximated using (5.53). From the point of view of numerical analysis, however, the Chebyshev polynomial  $\{T_n(x)\}$  is the best choice, primarily because at any point  $x$  within the domain  $[-1, 1]$ , the function  $T_n(x)$  has the smallest maximum deviation from the true function  $f(x)$  to be approximated. This property, which is unique to Chebyshev polynomials, is known as the **minimax property**. In general, polynomials endowed with the minimax property are very difficult to find, but fortunately, the Chebyshev polynomials fall into this category and, moreover, are easy to compute.

To show the minimax property of Chebyshev polynomials, we have to be aware of two of their other properties. The first is a concise formula for  $T_n(x)$  that is an alternative to those based on the Rodrigues formula.

♠ **Concise formula for Chebyshev polynomials:**

$$T_n(x) = \cos(n \cos^{-1} x) \quad (n = 0, 1, \dots). \quad (5.54)$$

The derivation of (5.54) requires some lengthy calculations, so we put it in the next subsection (see Sect. 5.3.2). Equation (5.54) implies that each  $T_n(x)$  has  $n$  zeros in the interval  $[-1, 1]$ , which are located at the points

$$x = \cos \left[ \frac{\pi}{n} \left( k - \frac{1}{2} \right) \right] \quad (k = 1, 2, \dots, n). \quad (5.55)$$

In this same interval, there are  $n + 1$  extrema (maxima and minima), located at

$$x = \cos \left( \frac{\pi}{n} k \right) \quad (k = 0, 1, \dots, n).$$

Note that  $T_n(x) = 1$  at all of the maxima, whereas  $T_n(x) = -1$  at all of the minima. This feature of  $T_n$  is exactly what makes the Chebyshev polynomials so useful in polynomial approximation of functions

*Remark.* Equation (5.54) combined with trigonometric identities can yield explicit expressions for  $T_n(x)$ :

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_2(x) = 2x^2 - 1, \quad T_3(x) = 4x^3 - 3x, \dots,$$

and more generally,

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x) \quad (n \geq 1).$$

The last expression is a special case of the general recurrence formula (5.40) derived in Sect. 5.2.3.

The second property of Chebyshev polynomials to be noted is the discrete orthogonality relation described below. (The proof is given in Sect. 5.3.3.)

♠ **Discrete orthogonal relation:**

If  $x_k$  ( $k = 1, \dots, n$ ) are the  $n$  zeros of  $T_n(x)$  given by (5.55) and  $i, j < n$ , then

$$\sum_{k=1}^n T_i(x_k) T_j(x_k) = \begin{cases} 0, & i \neq j, \\ n/2, & i = j \neq 0, \\ n, & i = j = 0. \end{cases} \quad (5.56)$$

From (5.54) and (5.56), we obtain the following theorem:

♠ **Theorem:**

Suppose  $f(x)$  to be an arbitrary function in the interval  $[-1, 1]$  and define  $c_j$  ( $j = 1, \dots, n$ ) by

$$c_j = \frac{2}{N} \sum_{k=1}^n f(x_k) T_{j-1}(x_k), \quad (5.57)$$

where  $x_k$  is the  $k$ th zero of  $T_n(x)$  given by (5.55). We then have

$$f(x) = \sum_{k=1}^n c_k T_{k-1}(x) - \frac{c_1}{2} \quad \text{for all } x = x_k. \quad (5.58)$$

What is remarkable is the fact that for  $x = x_k$ , the finite sum in (5.58) is equal to  $f(x)$  *exactly*. For  $x \neq x_k$ , the sum in (5.58) just *approximates*  $f(x)$ ; nevertheless the error can be reduced by increasing the degree  $n$  of the sum. Moreover, for practical use, we can truncate the sum in (5.58) to a much lower degree, for even if we do so, the approximation (5.58) is sufficiently accurate over the whole interval  $[-1, 1]$ , not only at the zeros of  $T_n(x)$ . This is in contrast to the case of approximations based on other polynomials, where the degree of summation  $n$  should be taken as large as possible to obtain high accuracy. In fact, this truncation capability is the reason Chebyshev polynomial expansion is far better than the other choices.

To examine the above statement, let us suppose that  $n$  is so large that (5.58) is virtually a perfect approximation of  $f(x)$ . We then consider the truncated approximation

$$f(x) \simeq \sum_{k=1}^m c_k T_{k-1}(x) - \frac{c_1}{2} \quad \text{with } m \ll n, \quad (5.59)$$

where the coefficients  $c_k$  are given in (5.57). The difference between (5.58) and (5.59) is given by

$$\sum_{k=m+1}^n c_k T_{k-1}(x), \quad (5.60)$$

which can be no larger than the sum of the neglected  $c_k$ 's as the  $T_n(x)$ 's are all bounded between  $\pm 1$ .

Now we consider the magnitude of the sum (5.60). We know that in general the  $c_k$ 's decrease rapidly with  $k$ , which follows intuitively from the definition (5.57). Hence, the magnitude of (5.60) is dominated by the term  $c_{m+1}T_m(x)$ , which is much less than unity for all  $x \in [-1, 1]$ . In addition,  $c_{m+1}T_m(x)$  is an oscillatory function with  $m+1$  equal extrema distributed almost uniformly over the interval  $[-1, 1]$ . These two features of the dominant term  $c_{m+1}T_m(x)$  result in smooth spreading out of the error of the approximation (5.59). This context implies that the Chebyshev approximation (5.59) is very nearly the same as the minimax polynomial that has the smallest maximum deviation from the true function  $f(x)$ .

### 5.3.2 A Concise Representation

The aim here is to derive the alternative representation of Chebyshev polynomials given in (5.54):

$$T_n(x) = \cos [n \cos^{-1}(x)].$$

We know that Chebyshev polynomials satisfy the relation

$$(1-x^2)\frac{d^2}{dx^2}T_n(x) - x\frac{d}{dx}T_n(x) + n^2T_n(x) = 0,$$

which can be rewritten in the form

$$\frac{d}{dx} \left( \sqrt{1-x^2} \frac{d}{dx} T_n(x) \right) + \frac{n^2}{\sqrt{1-x^2}} T_n(x) = 0. \quad (5.61)$$

We now apply the following lemma:

**♠ Lemma:**

Let  $p(x)$  and  $q(x)$  be two positive, continuously differentiable functions that satisfy the differential equation

$$\frac{d}{dx} \left[ p(x) \frac{d}{dx} y(x) \right] + q(x)y(x) = 0. \quad (5.62)$$

If the product  $p(x)q(x)$  is nonincreasing (or nondecreasing), then the relative maxima of  $[y(x)]^2$  form a nondecreasing (nonincreasing) set.

(The proof of this lemma is outlined in Exercise 1.) We can see that if

$$p(x) = \sqrt{1-x^2} \quad \text{and} \quad q(x) = \frac{n^2}{\sqrt{1-x^2}},$$

(5.61) corresponds to (5.62), which implies that the product  $pq$  is constant. Thus, according to the lemma, all relative maxima of  $T_n^2(x)$  must assume the same value.

Now we seek a polynomial  $T_n(x)$  of degree  $n$  that satisfies the condition

$$T_n^2(x) = 1 \quad \text{whenever} \quad T_n'(x) = 0.$$

That is,  $T_n^2(x) = 1$  at all  $x$  where  $T_n^2(x)$  has a relative maximum equal to 1. Clearly at these points, both  $T_n^2(x) - 1$  and  $[T_n'(x)]^2$  have double zeros. Then the function

$$\frac{T_n^2(x) - 1}{[T_n'(x)]^2} \tag{5.63}$$

is a rational function and all the zeros of the denominator also occur in the numerator. If we compare the degree of the polynomials in the denominator and in the numerator, it follows that (5.63) is a quadratic, and without loss of generality we have

$$\frac{T_n^2(x) - 1}{[T_n'(x)]^2} = \alpha(x^2 - 1). \tag{5.64}$$

The constant  $\alpha$  can be determined by dividing both sides by  $x^2$  and letting  $x$  approach infinity. Then, inserting a polynomial of degree  $n$  for  $T_n(x)$ , we obtain

$$\frac{1}{n^2} = \alpha \quad \text{so that} \quad T_n(x) = \cos [n \cos^{-1} x + c],$$

which yields

$$\frac{x^2 - 1}{n^2} \left( \frac{dT_n}{dx} \right)^2 = T_n^2 - 1. \tag{5.65}$$

Equation (5.65) is a differential equation for  $T_n(x)$  that determines the explicit form of our desired  $T_n(x)$ . To solve it, we set

$$T_n(x) = \cos \theta, \quad x = \cos \phi,$$

where  $\theta$  and  $\phi$  are functions of  $x$ . We then have

$$T_n^2(x) - 1 = -\sin^2 \theta$$

and

$$\frac{d}{dx} T_n(x) = \left( \frac{d}{d\phi} \cos \theta \right) \frac{d\phi}{dx} = \frac{\sin \theta}{\sin \phi} \frac{d\theta}{d\phi}.$$

Substituting these in (5.64) yields

$$\left(\frac{d\theta}{d\phi}\right)^2 = n^2 \quad \text{so that} \quad \theta = \pm n\phi + c,$$

and we get

$$T_n(x) = \cos(n \cos^{-1} x + c).$$

To determine  $c$ , we note that

$$T_n^2(\pm 1) = 1 = \cos(c).$$

Hence,  $c = 0$  and we eventually obtain

$$T_n(x) = \cos(n \cos^{-1} x). \quad (5.66)$$

### 5.3.3 Discrete Orthogonality Relation

We close this section by proving the discrete orthogonality relation (5.56) for Chebyshev polynomials.

**Proof (of the discrete orthogonality relation):** Let  $x_k$  ( $k = 1, 2, \dots, n$ ) be the  $n$  zeros of  $T_n(x)$ , which is given by

$$x_k = \cos \left[ \frac{\pi}{n} \left( k - \frac{1}{2} \right) \right] \quad (k = 1, 2, \dots, n).$$

Then the value of  $T_\ell(x)$  at  $x = x_k$ , in which  $\ell < n$  is assumed, reads

$$T_\ell(x_k) = \cos[\ell \cos^{-1}(x_k)] = \cos \left[ \frac{\pi \ell}{n} \left( k - \frac{1}{2} \right) \right].$$

Using the trigonometric identity, we have for  $\ell, m < n$ ,

$$\begin{aligned} & T_\ell(x_k) T_m(x_k) \\ &= \frac{1}{2} \cos \left[ \frac{\pi(\ell + m)}{2n} (2k - 1) \right] + \frac{1}{2} \cos \left[ \frac{\pi(\ell - m)}{2n} (2k - 1) \right]. \end{aligned} \quad (5.67)$$

If  $\ell = m = 0$ , this equals 1 so that we obtain

$$\sum_{k=1}^n T_\ell(x_k)^2 = \sum_{k=1}^n 1 = n. \quad (5.68)$$

Otherwise, if  $\ell = m \neq 0$ , the second term in the last line of (5.67) equals 1/2 and we have

$$\begin{aligned} \sum_{k=1}^n T_\ell(x_k)^2 &= \frac{n}{2} + \frac{1}{2} \sum_{k=1}^n \cos \left[ \frac{\ell \pi}{n} (2k - 1) \right] \\ &= \frac{n}{2} + \frac{\sin(2\ell\pi)}{4 \sin(\ell\pi/n)} = \frac{n}{2}, \end{aligned} \quad (5.69)$$

where we used the equation (see Exercise 2)

$$\sum_{k=1}^n \cos(2k-1)x = \frac{\sin 2nx}{2 \sin x} \quad (\text{for } x \neq 0).$$

In a similar manner, for the case  $\ell \neq m$  we find that

$$\sum_{k=1}^n T_\ell(x_k) T_m(x_k) = 0. \quad (5.70)$$

Equations (5.68), (5.69), and (5.70) together are identical to the desired result given in (5.56). ♣

## Exercises

1. Prove the lemma associated with the differential equation (5.62).

**Solution:** The proof is based on the nondecreasing property of the function defined by

$$f(x) = [y(x)]^2 + \frac{[p(x)y'(x)]^2}{p(x)q(x)},$$

in which the functions  $y(x)$ ,  $p(x)$ , and  $q(x)$  are assumed to satisfy the differential equation (5.62). The nondecreasing property of  $f(x)$  is verified by seeing its derivative:

$$f'(x) = 2yy' + \frac{2py'}{pq}(py')' + \left(\frac{1}{pq}\right)'(py')^2 = -\frac{(pq)'}{(pq)^2}(py')^2,$$

where we used the condition (5.62). From hypothesis,  $pq$  is nonincreasing, which implies  $(pq)' \leq 0$ . Hence, it is readily seen that  $f' \geq 0$ , i.e., that  $f$  is nondecreasing.

Now we realize that,  $y'$  must vanish wherever  $y(x)^2$  has a relative maximum so that  $f(x) = y^2$ . Suppose that  $x_1$  and  $x_2$  are two successive zeros of  $y'$ , such that  $x_1 < x_2$ . Since  $f(x)$  is nondecreasing, we have  $f(x_2) \geq f(x_1)$ , or equivalently,  $y^2(x_2) \geq y^2(x_1)$ , which means that the relative maxima of  $y^2$  form a nondecreasing set. This completes the proof of the lemma. ♣

2. Prove that  $\sum_{k=1}^n \cos(2k-1)x = \frac{\sin 2nx}{2 \sin x}$  (for  $x \neq 0$ ).

**Solution:** This equation is obtained by considering the sum

$$\begin{aligned} \sum_{k=1}^N e^{i(2k-1)x} &= e^{-ix} \sum_{k=1}^N e^{2ikx} = e^{-ix} \left[ \frac{1 - e^{2i(N+1)x}}{1 - e^{2ix}} - 1 \right] \\ &= e^{i(N-1)x} \cdot \frac{\sin(N+1)x}{\sin x} - e^{-ix}. \end{aligned}$$

Taking the real part of both sides yields

$$\begin{aligned}\sum_{k=1}^N \cos(2k-1)x &= \cos(N-1)x \cdot \frac{\sin(N+1)x}{\sin x} - \cos x \\ &= \frac{\sin 2Nx + \sin 2x}{2 \sin x} - \frac{2 \sin x \cos x}{2 \sin x} = \frac{\sin 2Nx}{2 \sin x}. \quad \clubsuit\end{aligned}$$

**3.** Derive the formula for Chebyshev polynomials:

$$\frac{1-t^2}{1-2tx+t^2} = T_0(x) + 2 \sum_{m=1}^{\infty} T_m(x)t^m,$$

where  $|t| < 1$ . Then, using this equation, prove that

$$\int_0^{2\pi} \frac{\cos \theta}{1-2t \cos \theta + t^2} d\theta = \frac{2\pi t^n}{1-t^2},$$

where  $n \geq 0$ .

**Solution:** It follows that

$$\begin{aligned}1 + \sum_{m=1}^{\infty} 2t^m \cos m\theta &= -1 + 2\operatorname{Re} \sum_{m=0}^{\infty} e^{im\theta} t^m = -1 + 2\operatorname{Re} 1/(1-te^{i\theta}) \\ &= (1-t^2)/(1-2tx+t^2),\end{aligned}$$

which the desired result. The next equation is found in the Fourier cosine series, where the coefficients can be obtained from

$$b_n = \frac{1}{\pi} \int_0^{2\pi} \frac{1-t^2}{1-2tx+t^2} \cos n\theta d\theta = 2t^n. \quad \clubsuit$$

## 5.4 Applications in Physics and Engineering

### 5.4.1 Quantum-Mechanical State in an Harmonic Potential

We now consider the application of **Hermite polynomials**  $H_n(x)$  to physical systems in the theory of **quantum mechanics**. We know that  $H_n(x)$  satisfies the following second-order differential equation:

$$H_n''(x) - xH_n'(x) + nH_n(x) = 0.$$

Let us introduce the related function

$$U_n(x) = e^{-x^2/4} H_n(x). \quad (5.71)$$

A simple calculation shows that

$$U_n''(x) + \left(n + \frac{1}{2} - \frac{x^2}{4}\right) U_n(x) = 0. \quad (5.72)$$

This equation is similar in form to the **Schrödinger equation** for a quantum particle whose motion is confined to an harmonic potential well. In fact, the Schrödinger equation is given by

$$\psi''(x) + \left(E - \frac{x^2}{2}\right) \psi(x) = 0, \quad (5.73)$$

where  $\psi(x)$  is the quantum wave function whose squared value at the position  $x = a$ , namely,  $|\psi(a)|^2$ , represents the **probability density** of the quantum particle being observed at  $x = a$ . The similarity between (5.72) and (5.73) implies that the product of the function defined by (5.71), i.e.,  $H_n(x)$ , and  $e^{-x^2/4}$  behaves as a wave function that describes the quantum particle in the potential well.

However, it should be noted that solutions of (5.73) do not always satisfy the condition

$$\int_{-\infty}^{\infty} |\psi(x)|^2 dx < \infty, \quad (5.74)$$

which must be satisfied for the solutions to be physically meaningful. By comparing (5.73) with (5.72), we see that whenever

$$E = E_n = 2n + 1, \quad (5.75)$$

we have

$$\psi_n(x) = c_n e^{-x^2/2} H_n(\sqrt{2}x),$$

which clearly satisfies the condition (5.74) if the constants  $c_n$  are chosen appropriately. Furthermore, the **uniqueness theorem** for solutions of ordinary differential equations (see Sect. 15.2.4) guarantees that the values of  $E$  given in (5.75) are the only ones for which (5.73) has solutions satisfying (5.75). These specific values of  $E$  are called the **eigenenergies** of the system, and the corresponding solutions  $\psi_n(x)$  are called **eigenfunctions**.

#### 5.4.2 Electrostatic potential generated by a multipole

Next, we briefly discuss the use of **Legendre polynomials** in describing the electrostatic potential field generated by a **multipole**. For simplicity, we first consider an electric dipole, i.e., a pair of positive and negative charges separated by an infinitesimal distance  $h$ . We choose our coordinate system such that both charges are located on the  $x$ -axis with the negative charge at the origin. The magnitude of the charges is taken to be  $\pm(1/h)$ . Then, the electrostatic potential field  $\Phi_2(P)$  with respect to a point  $P$  on the sphere  $x^2 + y^2 + z^2 = r^2$  is represented as

$$\begin{aligned}\Phi_2(P) &= \lim_{h \rightarrow 0} \frac{1}{h} \left( \frac{1}{\sqrt{(x-h)^2 + y^2 + z^2}} - \frac{1}{\sqrt{x^2 + y^2 + z^2}} \right) \\ &= \frac{\partial}{\partial x} \left( \frac{1}{r} \right) = -\frac{x}{r^3}.\end{aligned}$$

Therefore, when  $r = 1$ , we have

$$\Phi_2(P)|_{r=1} = -x = -P_1(x) \cdot 1!,$$

where  $P_n(x)$  is a Legendre polynomial.

Similar descriptions can be presented for high-degree multipoles. The potential  $\Phi_4(P)$  of a quadrupole is determined as follows: Consider a double negative charge  $-(2/h^2)$  located at the origin and two positive charges  $1/h^2$  located at the points  $(x, y, z) = (\pm h, 0, 0)$ . Then, the associated potential  $\Phi_4(P)$  at a point on a sphere of radius  $r$  is given by

$$\begin{aligned}\Phi_4(P) &= \lim_{h \rightarrow 0} \frac{1}{h^2} \left( \frac{1}{\sqrt{(x+h)^2 + y^2 + z^2}} - \frac{2}{\sqrt{x^2 + y^2 + z^2}} + \frac{1}{\sqrt{(x-h)^2 + y^2 + z^2}} \right) \\ &= \frac{\partial^2}{\partial x^2} \left( \frac{1}{r} \right) = -\frac{r^2 - 3x^2}{r^5},\end{aligned}$$

so for  $r = 1$ ,

$$\Phi_4(P)|_{r=1} = -1 + 3x^2 = P_2(x) \cdot 2!.$$

Similarly, for an octapole, we get

$$\Phi_8(P)|_{r=1} = \frac{\partial^3}{\partial x^3} \left( \frac{1}{r} \right) \Big|_{r=1} = -15x^3 + 9x = -P_3(x) \cdot 3!,$$

and in general

$$\Phi_{2^n}(P)|_{r=1} = \frac{\partial^n}{\partial x^n} \left( \frac{1}{r} \right) \Big|_{r=1} = (-1)^n P_n(x) \cdot n!.$$

The final result tells us that the potential of a  $2^n$ -pole is described by the product of the Legendre polynomial  $P_n(x)$  and the factor  $(-1)^n \cdot n!$ . By solving the previous equation for  $P_n(x)$ , we obtain the following expression for the  $n$ th Legendre polynomial:

$$P_n(x) = \frac{(-1)^n}{n!} \cdot \frac{\partial^n}{\partial x^n} \left( \frac{1}{r} \right) \Big|_{r=1}.$$



---

## Lebesgue Integrals

**Abstract** The concept of “measure” (Sect. 6.1.2) is important for an understanding of the theory of the Lebesgue integral. A measure is a generalization of the concept of length that allows us to quantify the length of a set that is composed of, for instance, an infinite number of infinitesimal points with a highly discontinuous distribution. Thus, the Lebesgue integral is an effective tool for integrating highly discontinuous functions that cannot be integrated using conventional Riemann integrals.

### 6.1 Measure and Summability

#### 6.1.1 Riemann Integral Revisited

It is certain that the **Riemann integral** is adequate for practical applications to most problems in physics and engineering, as the functions that we usually encounter are continuous (piecewise, at least) so that they are integrable by the Riemann procedure. In advanced subjects in mathematical physics, however, we come to a class of *highly irregular* functions where the concept of an ordinary Riemann integral is not applicable. In order to treat such functions, we have to employ another, more flexible integral than the Riemann integral. In this chapter, we present a concise description of the **Lebesgue integral**. The Lebesgue integral not only overcomes many of the difficulties inherent in the use of the Riemann integral, but its study has also generated new concepts and techniques that are extremely valuable in practical problems in modern physics and engineering.

At first, the cultivation of an intuitive feeling for the Lebesgue integral as an adjunct to formal manipulations and calculations is important, and we achieve this by comparing it with the Riemann integral. When defining the Riemann integral of a function  $f(x)$  on an interval  $I = [a, b]$ , we divide the entire interval  $[a, b]$  into small subintervals  $\Delta x_k = [x_k, x_{k+1}]$  such that

$$a = x_1 < x_2 < \cdots < x_{n+1} = b.$$

The finite set  $\{x_i\}$  of numbers is called a **partition**  $P$  of the interval  $I$ . Using this notation  $P$ , let us define, e.g., the sums

$$S_P(f) = \sum_{k=1}^n M_k(x_{k+1} - x_k), \quad s_P(f) = \sum_{k=1}^n m_k(x_{k+1} - x_k),$$

where  $M_k$  and  $m_k$  are the supremum and infimum of  $f(x)$  on the interval  $\Delta x_k = [x_k, x_{k+1}]$ , respectively, given by

$$M_k = \sup_{x \in \Delta x_k} f(x), \quad m_k = \inf_{x \in \Delta x_k} f(x). \quad (6.1)$$

Evidently, the relation  $S_P(f) \geq s_P(f)$  holds if the function  $f(x)$  is bounded on the interval  $I = [a, b]$ . We take the limit inferior (or limit superior) of the sums,

$$S(f) = \liminf_{n \rightarrow \infty} S_P, \quad s(f) = \limsup_{n \rightarrow \infty} s_P, \quad (6.2)$$

where all possible choices of the partition  $P$  are taken into account. The  $S(f)$  and  $s(f)$  are called the **upper** and **lower Riemann–Darboux integrals** of  $f$  over  $I$ , respectively. If the relation holds, i.e., if

$$S(f) = s(f) = A,$$

the common value  $A$  is called the **Riemann integral** and the function  $f(x)$  is called **Riemann integrable** such that

$$A = \int_a^b f(x) dx.$$

We note without proof that the following conditions ensure the existence of the Riemann integral of a function  $f(x)$ .

1.  $f(x)$  is continuous in  $I = [a, b]$ .
2.  $f(x)$  has only a finite number of discontinuities in  $I = [a, b]$ .

On the other hand, when the function  $f(x)$  exhibits too many points of discontinuity, the above definition is of no use in forming the integral. An illustrative example is given below.

*Examples* Assume an enumeration  $\{z_n\}$  ( $n = 1, 2, \dots$ ) of the rational numbers between 0 and 1 and let

$$f(x) = \begin{cases} 1 & (x = z_1, z_2, \dots, z_n) \\ 0 & \text{otherwise.} \end{cases}$$

That is, the function  $f(x)$  has the value unity if  $x$  is rational and the value zero if  $x$  is irrational. In any subdivision of the interval  $\Delta x_k \subset [0, 1]$ ,

$$m_k = 0, \quad M_k = 1,$$

and

$$s_P = 0, \quad S_P = 1.$$

Therefore, the upper and lower Darboux integrals are 1 and 0, respectively, whence  $f(x)$  has no Riemann integral.

### 6.1.2 Measure

The shortcoming of the Riemann procedure demonstrated above can be successfully overcome by employing Lebesgue's procedure. The latter requires a systematic way of assigning a **measure**  $\mu(X_i)$  to each subset of points  $X_i$ . In the remainder of this section, we learn about the basic properties of measure and its relevant materials, which serve as preliminaries to introduce the precise definition of Lebesgue integrals given in Sect. 6.2.

The measure for a subset of points is a generalization of the concepts of the length, area, and volume. Intuitively, it follows that the length of an interval  $[a, b]$  is  $b - a$ . Similarly, if we have two disjoint intervals  $[a_1, b_1]$  and  $[a_2, b_2]$ , it is natural to interpret the length of the set consisting of these two intervals as the sum  $(b_1 - a_1) + (b_2 - a_2)$ . However, the 'length' of a set of points of rational (or irrational) numbers on the line is not obvious. This context requires a rigorous mathematical definition of a measure of a point set, as shown below.

♠ **Measure of a set of points:**

A measure  $\mu(X)$  defined on a set of points  $X$  is a function with the following two properties:

1. If the set  $X$  is empty or consists of a single point,  $\mu(X) = 0$ ; otherwise,  $\mu(X) > 0$ .
2. The measure of the sum of two **nonoverlapping** sets is equal to the sum of the measures of these sets expressed by

$$\mu(X_1 + X_2) = \mu(X_1) + \mu(X_2) \quad \text{for } X_1 \cap X_2 = \emptyset. \quad (6.3)$$

In the above statement,  $X_1 + X_2$  denotes the set containing both elements of  $X_1$  and  $X_2$ , wherein each element is counted only once. If  $X_1$  and  $X_2$  overlap, (6.3) is replaced by

$$\mu(X_1 + X_2) = \mu(X_1) + \mu(X_2) - \mu(X_1 \cap X_2)$$

so that the points common to  $X_1$  and  $X_2$  will be counted only once.

Various kinds of measures have been thus far introduced in mathematics. Among them, is the following important example of measure that plays a central role in the subsequent discussions. Consider a monotonic increasing function  $\alpha(x)$  and let  $I$  be an interval (open or closed) with endpoints  $a$  and  $b$ . We define the  $\alpha$ -**measure** of  $I$  denoted by  $\mu_\alpha(I)$ , which takes different values depending on the types of endpoints  $a$  and  $b$  as shown below.

♠  **$\alpha$ -measure of intervals:**

$\alpha$ -measure of intervals are defined by

- $\mu_\alpha([a, b]) = \alpha(b^+) - \alpha(a^-)$  for the closed interval  $[a, b]$ ,
- $\mu_\alpha((a, b]) = \alpha(b^+) - \alpha(a^+)$  for the semiclosed interval  $(a, b]$ ,

- $\mu_\alpha([a, b)) = \alpha(b^-) - \alpha(a^-)$  for the semiclosed interval  $[a, b)$ ,
  - $\mu_\alpha((a, b)) = \alpha(b^-) - \alpha(a^+)$  for the open interval  $(a, b)$ ,
- where  $\alpha(a^-) = \lim_{\varepsilon \rightarrow 0} \alpha(a - \varepsilon)$  and  $\alpha(a^+) = \lim_{\varepsilon \rightarrow 0} \alpha(a + \varepsilon)$ .

By definition, the open interval  $(a, a)$  is an **empty set**, so that  $\mu_\alpha((a, a)) = 0$  for any  $a \in \mathbf{R}$ . The other cases of intervals  $(a, a]$  and  $[a, a)$  are also empty sets. Note that  $\mu_\alpha(I) \geq 0$  since  $\alpha(x)$  is a monotonically increasing function.

*Examples* Let  $\alpha(x)$  be the monotonically increasing function (see Fig. 6.1)

$$\alpha(x) = \begin{cases} 0, & x < 1, \\ \frac{1}{2}, & x = 1, \\ 1, & x > 1. \end{cases} \quad (6.4)$$

We then have

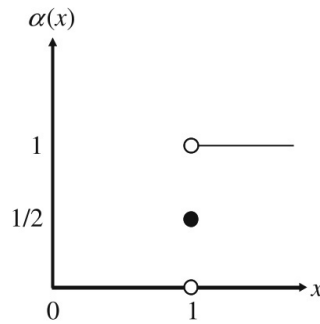
$$\mu_\alpha([0, 1)) = \alpha(1^-) - \alpha(0^-) = 0 - 0 = 0$$

and

$$\mu_\alpha([0, 1]) = \alpha(1^+) - \alpha(0^-) = 1 - 0 = 1.$$

Similarly,

$$\begin{aligned} \mu_\alpha([1, 2]) &= \mu_\alpha([1, 2)) = 2 - 0 = 2, \\ \mu_\alpha((1, 2]) &= \mu_\alpha((1, 2)) = 2 - 1 = 1. \end{aligned}$$



**Fig. 6.1.** The function  $\alpha(x)$  defined in (6.4)

### 6.1.3 The Probability Measure

The significance of measure is understood by illustrating the probability theory as an example. Probability theory deals with statistical properties of a

**random variable**  $x$  associated with an event occurring sequentially or simultaneously, where it is assumed that the average of  $x$  approaches a constant value as the number of observations increases.

Given a random variable  $x$ , its **expected** (or **mean**) **value** is defined by the integral

$$E\{x\} = \int_{-\infty}^{\infty} xp(x)dx, \quad (6.5)$$

where  $p(x) \geq 0$  is the **probability density function** of the random variable  $x$  defined by

$$p(x) = \frac{dP(x)}{dx},$$

with the **probability distribution function**  $P(x)$ . The function  $P(x)$  describes the probability that the event labeled  $x$  occurs. It follows intuitively that

$$P\{x_1 < x \leq x_2\} = \int_{x_1}^{x_2} p(x)dx \quad (6.6)$$

and

$$\int_{-\infty}^{\infty} p(x)dx = 1.$$

*Examples* For a discrete random variable  $\{x_i\}$ , the integral of (6.5) can be written as a sum:

$$E\{x\} = \sum_i x_i p_i.$$

In an experiment with dice, e.g., the probability of each event is given by

$$p_1 = p_2 = \cdots = p_6 = \frac{1}{6},$$

which yields

$$E\{x_i\} = \sum_{i=1}^6 x_i p_i = \frac{7}{3}.$$

In probability theory, the probability distribution function  $P(x)$  plays the role of measure. Assume a set of continuous real numbers,  $X = \{x \leq a\}$  and let the function  $\alpha(a)$  be the probability that  $x$  has a value no greater than  $a$ . The function  $\alpha(a)$  then reads

$$\alpha(a) = P(x \leq a), \quad (6.7)$$

where  $\alpha(-\infty^+) = 0$  and  $\alpha(\infty^-) = 1$ . Note that  $\alpha(a)$  is a monotonically increasing function. We have as well

$$P\{x_1 < x \leq x_2\} = \alpha(x_2) - \alpha(x_1),$$

since

$$P\{x \leq x_2\} = P\{x \leq x_1\} + P\{x_1 < x \leq x_2\}.$$

Therefore, we see that the probability distribution function  $P(x \in I)$  corresponds to the  $\alpha$ -measure for any interval  $I$ , as expressed by

$$\mu_\alpha(I) = P(x \in I),$$

which behaves as  $0 \leq \mu_\alpha(I) \leq 1$  for any  $I$ .

*Remark.* The mean value (6.5) of a random variable  $x$  can be interpreted as a **Riemann–Stieltjes integral**, rather than as an ordinary Riemann integral. To see this, we observe that the Riemann integral (6.5) can be expressed by the Riemann sum as

$$\int_{-\infty}^{\infty} xp(x)dx = \sum_{k=-\infty}^{\infty} \xi_k p(\xi_k)(x_{k+1} - x_k), \quad (6.8)$$

where  $\xi_k$  is any point on  $\Delta x_k$ . Since  $p(x_k)(x_{k+1} - x_k) = \Delta P\{x_k < x \leq x_{k+1}\}$  from (6.46), the mean value is written in the form

$$E\{x\} = \int_{-\infty}^{\infty} x dP = \int_{-\infty}^{\infty} x d\mu(x), \quad (6.9)$$

which is called the Riemann–Stieltjes integral of  $x$  with respect to  $\mu(x)$ .

#### 6.1.4 Support and Area of a Step Function

What follows is an important concept that we use together with the concept of measure to introduce the definition of the Lebesgue integral. Let  $I_i$  be any interval, and suppose that the **step function**  $\theta(x)$  given by

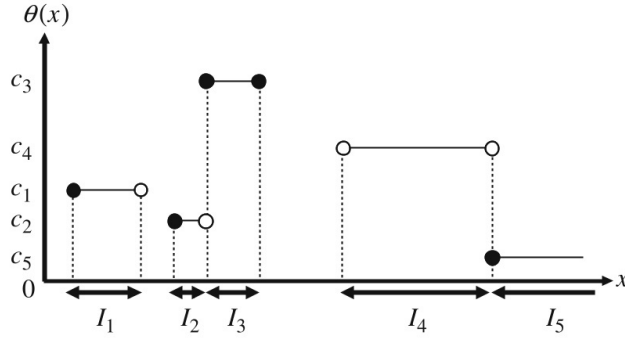
$$\theta(x) = \begin{cases} c_i, & x \in I_i, \quad i = 1, 2, \dots, n, \\ 0, & \text{otherwise,} \end{cases}$$

where a set  $\{c_1, c_2, \dots, c_n\}$  consists of finite and real numbers. We see that  $\theta$  is constant on each interval  $I_i$ , and zero elsewhere. We now introduce the following concept:

##### ♠ Support of a step function:

The disjoint set  $S = I_1 \cup I_2 \cup \dots \cup I_n \subseteq I$  on which  $\theta$  is nonzero is called the **support** of  $\theta(x)$ .

An example of the support  $\theta(x)$  is depicted in Fig. 6.2. When the support of a step function  $\theta$  has a *finite* total length, we associate it with the *area*  $A(\theta)$  between the graph of  $\theta$  and the  $x$ -axis, with the usual rule that areas below the  $x$ -axis have a *negative* sign. We refer to  $A(\theta)$  as the **area under the graph** of  $\theta$ .



**Fig. 6.2.** The disjoint set  $S = I_1 \cup I_2 \cup \cdots$  that serves as the support of  $\theta(x)$

Concepts such as support and area can apply to a linear combination of step functions. Suppose that  $\theta_1, \theta_2, \dots, \theta_n$  are step functions on the same interval  $I$ , all with supports of finite total length, and that  $a_1, a_2, \dots, a_n$  are finite real numbers. Then, the function  $\Theta(x)$  defined by

$$\Theta(x) = \sum_{j=1}^n a_j \theta_j(x) \quad \text{for } x \in I$$

is also a step function on  $I$ . The support of  $\Theta(x)$  has a finite length and the area under the graph of  $\Theta(x)$  is given by

$$A(\Theta) = \sum_{j=1}^n a_j A(\theta_j).$$

*Examples* Let  $\theta_1, \theta_2 : [0, 3) \rightarrow \mathbf{R}$  be defined by

$$\theta_1(x) = \begin{cases} 1 & \text{for } [0, 2), \\ 2 & \text{for } [2, 3), \end{cases} \quad \theta_2(x) = \begin{cases} -1 & \text{for } [0, 1], \\ 1 & \text{for } (1, 3). \end{cases} \quad (6.10)$$

Let  $\Theta = 2\theta_1 - \theta_2$ . Then

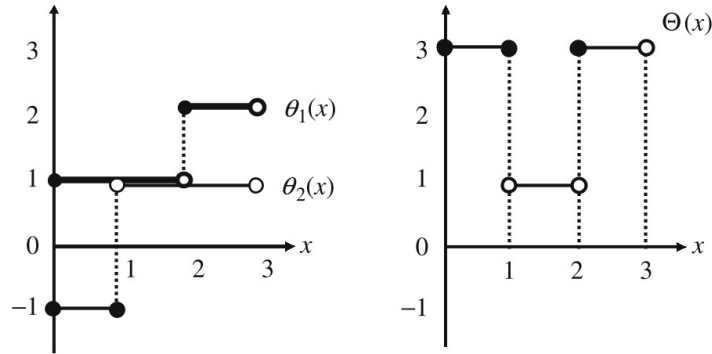
$$\Theta(x) = \begin{cases} 3 & \text{for } [0, 1], \\ 1 & \text{for } (1, 2), \\ 3 & \text{for } [2, 3). \end{cases} \quad (6.11)$$

These are plotted in Fig. 6.3. Clearly  $\Theta$  is a step function. Note also that the areas are

$$A(\theta_1) = 2(1) + 1(2) = 4, \quad A(\theta_2) = -1(1) + 2(1) = 1,$$

and

$$A(\Theta) = 1(3) + 1(1) + 1(3) = 7 = 2A(\theta_1) - A(\theta_2).$$



**Fig. 6.3.** The functions  $\theta_1(x)$ ,  $\theta_2(x)$ ,  $\Theta(x)$  given in (6.10) and (6.11), respectively

### 6.1.5 $\alpha$ -Summability

Now, we combine the concepts of  $\alpha$ -measure and support of a step function. Let  $\alpha(x)$  be a monotonically increasing function,  $I$  be any interval, and  $\theta(x)$  be a step function. We further assume that the support of  $\theta$  is a **simple set**, i.e., the union of a finite collection of disjoint intervals. For example, the set  $S = \bigcup_{k=1}^n I_k$  is a simple set if  $I_1, I_2, \dots, I_n$  are disjoint intervals. Then, the  **$\alpha$ -measure** of  $S$  is given by

$$\mu_\alpha(S) = \sum_{k=1}^n \mu_\alpha(I_k).$$

Observe that the value of  $\mu_\alpha(S)$  is independent of the way in which the set  $S$  is subdivided. Note also that

- (i)  $\mu_\alpha(S) \geq 0$  for any simple set  $S$ , and
- (ii) if  $S$  and  $T$  are simple sets such that  $S \subseteq T$ , then  $\mu_\alpha(S) \leq \mu_\alpha(T)$ .

We are now ready to present the following statement:

**♠  $\alpha$ -summability:**

A step function  $\theta(x)$  is  **$\alpha$ -summable** if the support of  $\theta$  has a finite  $\alpha$ -measure with respect to a given monotonically increasing function  $\alpha(x)$ .

Given an  $\alpha$ -summable step function  $\theta(x)$ , we associate it with a real number  $A_\alpha(\theta)$  defined by

$$A_\alpha(\theta) = \sum_{k=1}^n c_k \mu_\alpha(I_k), \quad (6.12)$$

where  $c_k$  is the amplitude of step function  $\theta(x)$  for  $x \in I_k$ . In general,  $A_\alpha(\theta)$  can be thought of as a generalized *area*. For example, when setting  $\alpha(x) = x$ , the measure  $\mu_\alpha(I_k)$  turns out to be just the ordinary length of the interval  $I_k$ , then  $A_\alpha(\theta)$  is just the area  $A(\theta)$  under the graph of  $\theta_j$  as defined in Sect. 6.1.4. However, if  $\alpha(x)$  has a more complicated function form, we get a different value of  $A_\alpha(\theta)$  from the above since in that case a *length* along the  $x$ -axis should be measured by the  $\alpha$ -measure rather than by ordinary length. An example of an actual calculation of  $A_\alpha(\theta)$  is provided in Exercise 2.

**Remark.** We shall see in Sect. 6.2.2 that the Lebesgue integral is defined by the limit  $n \rightarrow \infty$  of the sum in (6.12).

### 6.1.6 Properties of $\alpha$ -summable functions

We list some basic properties of  $\alpha$ -summable step functions without proof.

- If  $\theta(x)$  is a nonnegative  $\alpha$ -summable step function with respect to a given  $\alpha(x)$ , then  $A_\alpha(\theta) \geq 0$  and  $A_\alpha(0) = 0$ .
- If  $\theta_1$  and  $\theta_2$  are  $\alpha$ -summable step functions on the same interval  $I$  such that  $\theta_1 \leq \theta_2$  on  $I$ , then  $A_\alpha(\theta_1) \leq A_\alpha(\theta_2)$ .
- Let a set  $\{\theta_m\}$  be  $\alpha$ -summable step functions on the same interval  $I$ , and let  $\{a_m\}$  be finite real numbers. By defining  $\theta : I \rightarrow \mathbf{R}$  as

$$\theta(x) = \sum_{j=1}^m a_j \theta_j(x)$$

for all  $x \in I$  ( $\theta$  is also an  $\alpha$ -summable step function on  $I$ ), we have

$$A_\alpha(\theta) = \sum_{j=1}^m a_j A_\alpha(\theta_j).$$

## Exercises

1. Assume a monotonically increasing function  $\alpha(x)$  defined by

$$\alpha(x) = \begin{cases} 0, & x \in (-\infty, 1), \\ x^2 - 2x + 2, & x \in [1, 2), \\ 3, & x = 2, \\ x + 2, & x \in (2, \infty). \end{cases}$$

Calculate  $A_\alpha(\theta)$  for each of the two step functions:

$$\theta_1(x) = \begin{cases} -1, & x \in [0, 1), \\ 2, & x \in [1, 3], \end{cases}$$

and

$$\theta_2(x) = \begin{cases} -1, & x \in [0, 1], \\ 2, & x \in (1, 3]. \end{cases}$$

**Solution:** Since

$$\mu_\alpha([0, 1)) = \alpha(1^-) - \alpha(0^-) = 0 - 0 = 0,$$

$$\mu_\alpha([1, 3]) = \alpha(3^+) - \alpha(1^-) = 5 - 0 = 5,$$

we have

$$A_\alpha(\theta_1) = (-1)0 + 2(5) = 10.$$

For  $\theta_2$ , on the other hand, we have a different result since

$$\mu_\alpha([0, 1]) = \alpha(1^+) - \alpha(0^-) = 1 - 0 = 1,$$

$$\mu_\alpha((1, 3]) = \alpha(3^+) - \alpha(1^+) = 5 - 1 = 4,$$

which yields

$$A_\alpha(\theta_2) = (-1)1 + 2(4) = 7.$$

It is noteworthy that the values of  $A_\alpha(\theta_1)$  and  $A_\alpha(\theta_2)$  are different, although the area  $A(\theta)$  for them is the same. The difference comes from the fact that  $\alpha$  has a discontinuity at the single point where  $\theta_1$  and  $\theta_2$  have different values. ♣

**2.** Evaluate  $A_\alpha(\theta)$  of the step function:

$$\theta(x) = \begin{cases} 2, & x \in (-\infty, 0], \\ 1, & x \in (0, \infty), \end{cases}$$

which is associated with the  $\alpha$ -measure:

$$\alpha(x) = \begin{cases} 0, & x < 0, \\ \frac{1}{2}, & x = 0, \\ 1, & x > 0. \end{cases}$$

**Solution:** Since

$$\mu_\alpha((-\infty, 0]) = \alpha(0^+) - \alpha(-\infty^+) = \frac{1}{2} - 0 = \frac{1}{2},$$

$$\mu_\alpha((0, \infty)) = \alpha(\infty^-) - \alpha(0^+) = 1 - \frac{1}{2} = \frac{1}{2},$$

we have

$$A_\alpha(\theta_1) = 1 \left( \frac{1}{2} \right) + 2 \left( \frac{1}{2} \right) = \frac{3}{2}. \quad \clubsuit$$

3. Show that the function

$$f(x) = \lim_{n \rightarrow \infty} \lim_{m \rightarrow \infty} (\cos 2\pi m!x)^n,$$

called **Dirichlet's function**, takes the form

$$f(x) = \begin{cases} 1 & \text{for all rational numbers } x, \\ 0 & \text{otherwise.} \end{cases}$$

**Solution:** When  $x$  is a rational number, it is expressed by a fraction  $p/q$  with relatively prime integers  $p$  and  $q$ . Hence, for sufficiently large  $m$ , the product  $m!x$  becomes an integer since

$$m!x = m \cdot (m-1) \cdots (q+1) \cdot p \cdot (q-1) \cdots 2 \cdot 1.$$

Thus we have  $\cos 2\pi m!x = 1$ . Otherwise, if  $x$  is an irrational number,  $m!x$  is also an irrational for any  $m$ , so that  $|\cos 2\pi m!x| < 1$ . As a result, we obtain

$$\lim_{n \rightarrow \infty} \lim_{m \rightarrow \infty} (\cos 2\pi m!x)^n = \begin{cases} 1 : x \text{ is a rational,} \\ 0 : x \text{ is an irrational.} \end{cases} \quad \clubsuit$$

## 6.2 Lebesgue Integral

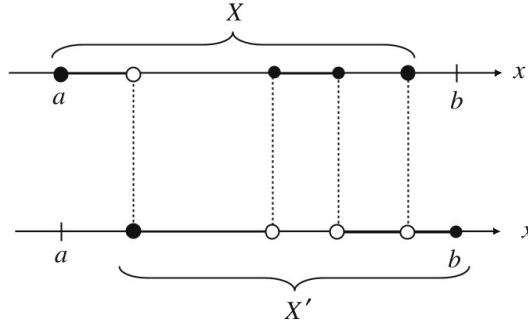
### 6.2.1 Lebesgue Measure

The Lebesgue integral procedure essentially reduces to *finding a measure for sets of arguments*. In particular if a set consists of too many points of discontinuity, we need a way to define its measure that is known as the **Lebesgue measure**. In this subsection, we explain how to construct the Lebesgue measure of a point set.

As a simple example, let us consider a finite interval  $[a, b]$  of length  $L$ . This can be decomposed into two sets: a set  $X$  consisting of some of the points  $x \in [a, b]$  and its **complementary set**  $X'$  consisting of all points  $x \in [a, b]$  that do not belong to  $X$ . A schematic view of  $X$  and  $X'$  is shown in Fig. 6.4. Both  $X$  and  $X'$  may be sets of several continuous line segments or sets of **isolated points**.

We would like to evaluate the measure of  $X$ . To do this, we cover the set of points  $X$  by nonoverlapping intervals  $A_i \subset [a, b]$  such as

$$X \subset (A_1 + A_2 + \cdots).$$



**Fig. 6.4.** A set  $X$  and its complementary set  $X'$

If we denote the length of  $\Lambda_k$  by  $\ell_k$ , the sum of  $\ell_k$  must satisfy the inequality

$$0 \leq \sum_k \ell_k \leq L.$$

In particular, the smallest value of the sum  $\sum_i \ell_i$  is referred to as the **outer measure** of  $X$  and is denoted by

$$\mu_{\text{out}}(X) = \inf \left( \sum_k \ell_k \right).$$

In the same manner, we can find intervals  $\Lambda'_k \subset [a, b]$  of lengths  $\ell'_1, \ell'_2, \dots$  that cover the complementary set  $X'$  such that

$$X' \subset (\Lambda'_1 + \Lambda'_2 + \dots), \quad 0 \leq \sum_k \ell'_k \leq L.$$

Here we define another kind of measure denoted by

$$\mu_{\text{in}}(X) \equiv L - \mu_{\text{out}}(X') = L - \inf \left( \sum_k \ell'_k \right), \quad (6.13)$$

which is called the **inner measure** of  $X$ . Note that the inner measure of  $X$  is defined by the outer measure of  $X'$ , not of  $X$ . It is a straightforward matter to prove the inequality

$$0 \leq \mu_{\text{in}}(X) \leq \mu_{\text{out}}(X). \quad (6.14)$$

Specifically, if

$$\mu_{\text{in}}(X) = \mu_{\text{out}}(X),$$

it is called the **Lebesgue measure** of the point set  $X$ , denoted by  $\mu(X)$ . Clearly, when  $X$  contains all the points of  $[a, b]$ , the smallest interval that covers  $[a, b]$  is  $[a, b]$  itself, and thus  $\mu(X) = L$ .

Our results are summarized below.

♠ **Lebesgue measure:**

A set of points  $X$  is said to be **measurable** with the **Lebesgue measure**  $\mu(X)$  if and only if  $\mu_{\text{in}}(X) = \mu_{\text{out}}(X) \equiv \mu(X)$ .

*Remark.* An unbounded point set  $X$  is measurable if and only if  $(-c, c) \cap X$  is measurable for all  $c > 0$ . In this case, we define  $\mu(X) = \lim_{c \rightarrow \infty} \mu[(-c, c) \cap X]$ , which may or may not be finite.

### 6.2.2 Definition of the Lebesgue Integral

We are now in a position to define the **Lebesgue integral**. Let the function  $f(x)$  be defined on a set  $X$  that is bounded:

$$0 \leq f_{\min} \leq f(x) \leq f_{\max}.$$

We partition the ordinate axis by the sequence  $\{f_k\}$  ( $1 \leq k \leq n$ ) so that  $f_1 = f_{\min}$  and  $f_n = f_{\max}$ . Owing to the one-to-one correspondence between  $x$  and  $f(x)$ , there should exist sets  $X_i$  of values  $x$  such that

$$f_k \leq f(x) < f_{k+1} \quad \text{for } x \in X_k \quad (1 \leq k \leq n-1), \quad (6.15)$$

as well as a set  $X_n$  of values  $x$  such that  $f(x) = f_n$ . Each set  $X_k$  assumes a measure  $\mu(X_k)$ . Thus we form the sum of products  $f_k \cdot \mu(X_k)$  of all possible values of  $f$ , called the **Lebesgue sum**:

$$\sum_{k=1}^n f_k \cdot \mu(X_k). \quad (6.16)$$

If the sum (6.16) converges to a finite value when taking the limit  $n \rightarrow \infty$  such that

$$\max |f_k - f_{k+1}| \rightarrow 0,$$

then the limiting value of the sum is called the **Lebesgue integral** of  $f(x)$  over the set  $X$ .

The formal definition of the Lebesgue integral is given below.

♠ **Lebesgue integral:**

Let  $f(x)$  be a nonnegative function defined on a **measurable set**  $X$  and divide  $X$  into a finite number of subsets such as

$$X = X_1 + X_2 + \cdots + X_n. \quad (6.17)$$

Let  $f_k = \inf_{x \in X_k} f(x)$  to form the sum

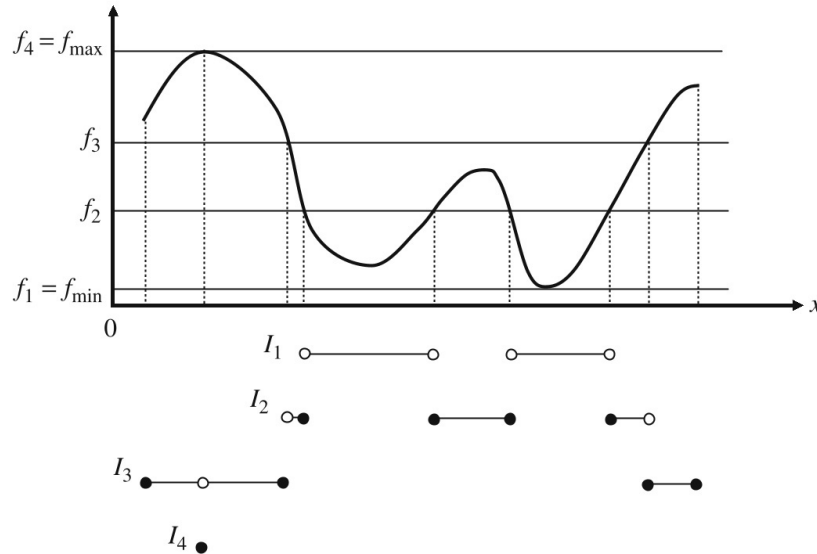
$$\sum_{k=1}^n f_k \mu(X_k). \quad (6.18)$$

Then the **Lebesgue integral** of  $f(x)$  on  $X$  is defined by

$$\int_X f d\mu \equiv \lim_{\max |f_k - f_{k-1}| \rightarrow 0} \left[ \sum_{k=1}^n f_k \mu(X_k) \right],$$

where all possible choices of partition (6.17) are considered.

Figure 6.5 is a schematic illustration of the Lebesgue procedure. Obviously, the value of the Lebesgue sum (6.16) depends on our choice of partition. If we take an alternative partition instead of (6.17), the value of the sum also changes. Among the infinite variety of choices, the partition that maximizes the sum (6.17) gives the Lebesgue integral of  $f(x)$ . That a function is **Lebesgue integrable** means that the limit superior of the sum (6.18) is determined independently of our choice of the partition of the  $x$ -axis.



**Fig. 6.5.** An illustration of the Lebesgue procedure

### 6.2.3 Riemann Integrals vs. Lebesgue Integrals

Before proceeding further with this discussion, we compare the definitions of Riemann and Lebesgue integrals for a better understanding of the significance of the latter. In the language of measure, the **Riemann integral** of a function

$f(x)$  defined on the set  $X$  is obtained by dividing  $X$  into nonoverlapping subsets  $X_i$  as

$$X = X_1 + X_2 + \cdots + X_n, \quad X_i \cap X_j = 0, \quad \text{for any } i, j,$$

followed by setting the **Riemann sum**

$$\sum_{k=1}^n f(\xi_k) \mu(X_k). \quad (6.19)$$

Here, the measure  $\mu(X_k)$  is identified with the length of the subset  $X_k$ , and  $\xi_k$  assumes any point that belongs to  $X_k$ . We increase the number of subsets  $n \rightarrow \infty$  such that

$$\mu(X_k) \rightarrow 0 \quad \text{for any } X_k,$$

and if the limit of the sum (6.19) exists and is independent of the subdivision process, it is called the **Riemann integral** of  $f(x)$  over  $X$ . Obviously, the Riemann integral can be defined under the condition that all values of  $f(x)$  defined over  $X_k$  tend to a common limit as  $\mu(X_k) \rightarrow 0$ . Such a requirement excludes any possibility of defining the Riemann integral for functions having too many points of discontinuity.

**Remark.** In view of the analogy between the sum (6.12) and (6.18), we may say that, in a sense, the Lebesgue integral is the limit  $n \rightarrow \infty$  of the quantity  $A_\alpha(\theta)$ .

Although the Lebesgue sum given in (6.16) is apparently similar to the Riemann sum given in (6.19), they are intrinsically different. In the Riemann sum (6.19),  $f(\xi_i)$  is the value of  $f(x)$  at an *arbitrary* point  $\xi_i \in X_i$ . Thus the value of  $\xi_i$  is allowed to vary within each subset, which causes an indefiniteness in the value of  $f(\xi_i)$  within each subset. On the other hand, in the Lebesgue sum (6.16), the value of  $f_i$  corresponding to each subset  $X_i$  has a definite value. Therefore, for the existence of the Lebesgue integral, we no longer need local smoothness of  $f(x)$ . As a result, the conditions imposed on the integrated function become very *weak* compared with the case of the Riemann integral.

#### 6.2.4 Properties of the Lebesgue Integrals

Several properties of the Lebesgue integral are given below without proof.

1. If  $f(x)$  is the Lebesgue integrable on  $X$  and if  $X = X_1 + X_2 + \cdots + X_n$ , then

$$\int_X f d\mu = \sum_{i=1}^n \int_{X_i} f d\mu.$$

2. If two functions  $f(x)$  and  $g(x)$  are both Lebesgue integrable on  $X$  and if  $f(x) \leq g(x)$  for any  $x \in X$ , then

$$\int_X f d\mu \leq \int_X g d\mu.$$

3. If  $\mu(X) = 0$ , then  $\int_X f(x) dx = 0$ .

4. If the integral  $\int_X f(x) dx$  is finite, then the subset of  $X$  defined by

$$X' = \{x \mid f(x) = \pm\infty\}$$

has zero measure. This means that in order for the integral to converge, the measure of a set of points  $x$  at which  $f(x)$  diverges is necessarily zero.

5. Suppose that  $\int_X f(x) dx$  is finite and that  $X' \subset X$ . If we make  $\mu(X') \rightarrow 0$ , then

$$\int_{X'} f d\mu \rightarrow 0.$$

6. When  $f(x)$  on  $X$  takes both positive and negative values, its Lebesgue integral is defined by

$$\int_X f d\mu = \int_X f^+ d\mu + \int_X f^- d\mu \quad (6.20)$$

and

$$\int_X |f| d\mu = \int_X f^+ d\mu - \int_X f^- d\mu, \quad (6.21)$$

where

$$f^+(x) = \begin{cases} f(x) & \text{for } \{x; f(x) \geq 0\}, \\ 0 & \text{for } \{x; f(x) < 0\}, \end{cases}$$

and

$$f^-(x) = \begin{cases} 0 & \text{for } \{x; f(x) \geq 0\}, \\ -f(x) & \text{for } \{x; f(x) < 0\}. \end{cases}$$

Definition (6.21) is justified except when both integrals on the right-hand side diverge.

### 6.2.5 Null-Measure Property of Countable Sets

Let us show that any countable set has a Lebesgue measure equal to zero. A rigorous definition of countable sets is given herewith.

#### ♠ Countable set:

A finite or infinite set  $X$  is **countable** (or **enumerable**) if and only if it is possible to establish a reciprocal one-to-one correspondence between its elements and the elements of a set of real integers.

It follows that every finite set is countable and that every subset of a countable set is also countable. Any countable set is associated with a specific number, called the **cardinal number**, defined below.

♠ **Cardinal numbers:**

Two sets  $X_1$  and  $X_2$  are said to have the same cardinal number if and only if there exists a reciprocal one-to-one correspondence between their respective elements.

**Remark.** A set  $X$  is called an **infinite set** if it has the same cardinal number as one of its subsets; otherwise,  $X$  is called a **finite set**.

It should be stressed that an infinite set may or may not be countable. When a given infinite set is countable, then its cardinal number is denoted by  $\aleph_0$ , which is the same as the cardinal number of the set of the positive real integers. Furthermore, the cardinal number of every noncountable set is denoted by  $\aleph$ , which is identified with the cardinal number of the set of all real numbers (or the set of points on a continuous line). Cardinal numbers of infinite sets,  $\aleph_0$  and  $\aleph$ , are called **transfinite numbers**.

The most important property of countable sets in terms of measure theory is given below.

♠ **Theorem:**

Any countable set (finite or infinite) has a Lebesgue measure of zero, namely, **null measure**.

*Examples* An illustrative example is the set of rational numbers that has measure zero as shown earlier. The countability of this set follows from the fact that it can be arranged in a sequence of proper fractions as

$$0, 1, \frac{1}{2}, \frac{1}{3}, \frac{2}{3}, \frac{1}{4}, \frac{3}{4}, \frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}, \dots$$

Accordingly, since the set of all rational numbers in the interval  $[0, 1]$  has zero measure, the Lebesgue integral of Dirichlet's function  $\chi(x)$  over this interval is well defined and equal to zero.

Another well-known example of the set of measure zero is the **Cantor set**, which is demonstrated in Exercise 2.

### 6.2.6 The Concept of Almost Everywhere

We have observed that sets of measure zero make no contribution to Lebesgue integrals. This fact provides a concept of an **equality almost everywhere**

for measurable functions, which plays an important role in developing the theory of function analysis.

♠ **Equality almost everywhere:**

Two functions  $f(x)$  and  $g(x)$  defined on the same set  $X$  are said to be equal almost everywhere with respect to a measure  $\mu(X)$  if

$$\mu\{x \in X; f(x) \neq g(x)\} = 0.$$

We extend this terminology to other circumstances as well. In general, a property is said to hold **almost everywhere** on  $X$  if it holds at all points of  $X$  except on a set of measure zero. Thus two functions  $f(x)$  and  $g(x)$  are said to be equivalent (written  $f \sim g$ ) if they coincide almost everywhere. For example, Dirichlet's function mentioned earlier is equivalent almost everywhere to the function  $g(x) \equiv 0$ .

Since the behavior of functions on sets of measure zero is often unimportant, it is natural to introduce the following generalization of the ordinary notion of the convergence of a sequence of functions:

♠ **Convergence almost everywhere:**

A sequence of functions  $\{f_n(x)\}$  defined on a set  $X$  is said to **converge almost everywhere** to a function  $f(x)$  if

$$\lim_{n \rightarrow \infty} f_n(x) = f(x) \tag{6.22}$$

for all  $x \in X$  except for points of measure zero.

*Examples* A typical example is the sequence

$$\{f_n(x)\} = \{(-x)^n\}$$

defined on  $[0, 1]$ . It converges almost everywhere to the function  $f(x) \equiv 0$ ; in fact it converges everywhere except at the point  $x = 1$ .

## Exercises

1. Show that the set of all rational numbers in the interval  $[0, 1]$  has a Lebesgue measure equal to zero.

**Solution:** Denote by  $X'$  the set of irrational numbers that is complementary to  $X$  and the entire interval  $[0, 1]$  by  $I$ . Since  $\mu(I) = 1$ , the outer measure of  $X'$  reads

$$\mu_{\text{out}}(X') = \mu_{\text{out}}(I - X) = \mu_{\text{out}}(I) - \mu_{\text{out}}(X) = 1 - \mu_{\text{out}}(X).$$

By definition, the inner measure of  $X$  is given by

$$\mu_{\text{in}}(X) = \mu_{\text{in}}(I) - \mu_{\text{out}}(X') = 1 - [1 - \mu_{\text{out}}(X)] = \mu_{\text{out}}(X).$$

The last equality asserts that the set  $X$  is **Lebesgue measurable**). The remaining task is to evaluate the value of  $\mu(X) = 0$ .

Let  $x_k$  ( $k = 1, 2, \dots, n, \dots$ ) denote the points of rational numbers in the interval  $I$ . We cover each point  $x_1, x_2, \dots, x_n, \dots$  by an open interval of length  $\varepsilon/2, \varepsilon/2^2, \dots, \varepsilon/2^n, \dots$ , respectively, where  $\varepsilon$  is an arbitrary positive number. Since these intervals may overlap, the entire set can be covered by an open set of measure not greater than

$$\sum_{n=1}^{\infty} \frac{\varepsilon}{2^n} = \frac{\varepsilon}{2(1 - \frac{1}{2})} = \varepsilon.$$

Since  $\varepsilon$  can be made arbitrarily small, we find that  $\mu_{\text{out}}(X) = 0$ .

Hence, from (6.18) we immediately have  $\mu(X) = 0$ . ♣

- 2.** Evaluate the measure of a **Cantor set**, an infinite set constructed as follows: **(i)** From the closed interval  $[0, 1]$ , delete the open interval  $(1/3, 2/3)$  that forms its middle third; **(ii)** from each of the remaining intervals  $[0, 1/3]$  and  $[2/3, 1]$  delete the middle third; **(iii)** continue this process of deleting the middle thirds indefinitely to obtain the point set on the line that remains after all these open intervals.

**Solution:** Observe that at the  $k$ th step, we have thrown out  $2^{k-1}$  adjacent intervals of length  $1/3^k$ . Thus the sum of the lengths of the intervals removed is equal to

$$\frac{1}{3} + \frac{2}{9} + \frac{4}{27} + \dots + \frac{2^{n-1}}{3^n} + \dots = \lim_{n \rightarrow \infty} \frac{\frac{1}{3} [1 - (\frac{2}{3})^n]}{1 - \frac{2}{3}} = 1.$$

This is just the measure of the open set  $P'$  that is the **complement** of  $P$ . Therefore, the Cantor set  $P$  itself has null measure

$$\mu(P) = 1 - \mu(P') = 1 - 1 = 0. \quad \clubsuit$$

- 3.** Show that if  $f(x)$  is nonnegative and integrable on  $X$ , then

$$\mu[x \in X, f(x) \geq c] \leq \frac{1}{c} \int_X f d\mu,$$

which is known as, **Chebyshev's inequality**.

**Solution:** Set  $X' = \{x \in X, f(x) \geq c\}$  to observe that

$$\int_X f d\mu = \int_{X'} f d\mu + \int_{X-X'} f d\mu \geq \int_{X'} f d\mu \geq c\mu(X'). \quad \clubsuit$$

4. Show that if  $\int_X |f| d\mu = 0$ , then  $f(x) = 0$  **almost everywhere**.

**Solution:** By Chebyshev's inequality,

$$\mu \left[ x \in X, |f(x)| \geq \frac{1}{n} \right] \leq n \int_X |f| d\mu = 0$$

for all  $n = 1, 2, \dots$ . Therefore, we have

$$\mu [x \in X, f(x) \neq 0] \leq \sum_{n=1}^{\infty} \mu \left[ x \in X, |f(x)| \geq \frac{1}{n} \right] = 0. \quad \clubsuit$$

## 6.3 Important Theorems for Lebesgue Integrals

### 6.3.1 Monotone Convergence Theorem

Our current task is to examine whether or not the equality

$$\lim_{n \rightarrow \infty} \int_a^b f_n(x) dx = \int_a^b f(x) dx \quad (6.23)$$

is valid under the Lebesgue procedure. This problem can be clarified by referring to two important theorems concerning the convergence property of Lebesgue integrals; the **monotone convergence theorem** and the **dominated convergence theorem**. Neither theorem is valid if we restrict our attention to Riemann integrable functions. We observe that, owing to the two convergence theorems, Lebesgue theory offers a considerable improvement over Riemann theory with regard to convergence properties.

In what follows, we assume that  $X$  is a set of real numbers, and that  $\{f_n\}$  is a sequence of functions defined on  $X$ .

#### ♠ Monotone convergence theorem:

If  $(f_n)$  is a sequence such that  $0 \leq f_n \leq f_{n+1}$  for all  $n \geq 1$  in  $X$  and  $f = \lim_{n \rightarrow \infty} f_n$ , then

$$\lim_{n \rightarrow \infty} \int_X f_n d\mu = \int_X \lim_{n \rightarrow \infty} f_n d\mu = \int_X f d\mu.$$

**Remark.** The monotone convergence theorem states that in the case of Lebesgue integrals, the conditions to reverse the order of limit and integration are much weaker than in the case of Riemann integrals; i.e., only the **point-wise convergence** of  $f_n(x)$  to  $f(x)$  is required in the Lebesgue case, whereas in the Riemann case we must have **uniform convergence** of  $f_n(x)$  to  $f(x)$ .

**Proof (of the monotone convergence theorem):** The hypothesis  $0 \leq f_n \leq f_{n+1}$  implies that

$$0 \leq \int_X f_n d\mu \leq \int_X f_{n+1} d\mu,$$

which indicates that the sequence  $\{\int_X f_n d\mu\}$  increases monotonically with respect to  $n$ ; thus its limit  $n \rightarrow \infty$  exists as we denote it by  $M$  (possibly equal to  $\infty$ ). In addition, by hypothesis

$$\int_X f_n d\mu \leq \int_X f d\mu, \quad \text{for all } n. \quad (6.24)$$

Since (6.24) is true for arbitrary  $n$ , we have

$$M = \lim_{n \rightarrow \infty} \left[ \int_X f_n d\mu \right] \leq \int_X f d\mu.$$

Therefore, if we can verify the opposite inequality

$$M \geq \int_X f d\mu, \quad (6.25)$$

we will get the desired result,

$$M = \lim_{n \rightarrow \infty} \int_X f_n d\mu = \int_X \lim_{n \rightarrow \infty} f_n d\mu = \int_X f d\mu.$$

To show (6.25), let  $c$  be a number such that  $c \in (0, 1)$  and introduce the point set

$$X_n = \{x : cf(x) \leq f_n(x)\}.$$

Owing to the monotonically increasing property of the sequence  $\{f_n(x)\}$  with regard to  $n$ , the set  $X_n$  satisfies the inclusion relation

$$X_1 \subset X_2 \subset X_3 \subset \cdots \quad \text{and} \quad \bigcup_{n=1}^{\infty} X_n = X.$$

In addition, the increasing property of the sequence  $\{\int_{X_n} f_n d\mu\}$  yields

$$c \int_{X_n} f d\mu \leq \int_{X_n} f_n d\mu \leq \lim_{n \rightarrow \infty} \left[ \int_{X_n} f_n d\mu \right] = M. \quad (6.26)$$

Since (6.26) must hold for any  $n$ , we have

$$c \int_X f d\mu \leq M. \quad (6.27)$$

Furthermore, since (6.27) is true for all  $c \in (0, 1)$ , we have

$$\int_X f d\mu \leq M. \quad (6.28)$$

Note that the substitution  $c = 1$  into (6.27) is allowed because the symbol  $\leq$ , not  $<$ , is involved in (6.27). ♣

### 6.3.2 Dominated Convergence Theorem (I)

In the previous argument, we saw that the order of limit and integration can be reversed when considering monotonically increasing sequences of functions. In practice, however, the requirement in the monotone convergence theorem, i.e., that the sequence  $\{f_n(x)\}$  must be monotone increasing, is sometimes very inconvenient. In this subsection, we examine the same issue for more general sequences of functions, i.e., nonmonotone sequences satisfying some looser conditions and their limit passage. Our current objective is to prove the theorem below.

♠ **Dominated convergence theorem:**

Let  $\{f_n\}$  be a sequence of functions for almost everywhere on  $X$  such that **(a)**  $\lim_{n \rightarrow \infty} f_n(x) = f(x)$ , and **(b)** there exists a nonnegative  $g$  such that  $|f_n| \leq g$  for all  $n \geq 1$ . Then, we have

$$\lim_{n \rightarrow \infty} \int_X f_n d\mu = \int_X f d\mu.$$

*Remark.* Note that the condition imposed on the theorem above is that the sequences  $\{f_n\}$  should be bounded almost everywhere. This condition is clearly looser than that imposed in the monotone convergence theorem. Hence, the monotone convergence theorem can be regarded as a special case of the dominated convergence theorem.

### 6.3.3 Fatou Lemma

The proof of the dominated convergence theorem requires the lemma given below.

♠ **Fatou lemma:**

If  $f_n(x) \geq 0$  for all  $n$  and for almost everywhere in a bounded measurable set  $X$  and if  $\lim_{n \rightarrow \infty} f_n(x) = f(x)$ , then

$$\int_X \left[ \liminf_{n \rightarrow \infty} f_n \right] d\mu = \int_X f d\mu \leq \liminf_{n \rightarrow \infty} \left[ \int_X f_n d\mu \right],$$

where the definition is

$$\liminf_{n \rightarrow \infty} f_n = \lim_{n \rightarrow \infty} \left[ \inf_{k \geq n} f_k \right].$$

*Proof* Let  $g_n = \inf_{k \geq n} f_k$ . Since the sequence  $g_n(x)$  is nonnegative and non-decreasing, we have

$$\lim_{n \rightarrow \infty} g_n = \liminf_{n \rightarrow \infty} f_n.$$

(See Sect. 2.1.4 for the precise definition of  $\liminf$ .) In addition, the monotone convergence theorem implies that

$$\lim_{n \rightarrow \infty} \int_X g_n d\mu = \int_X \lim_{n \rightarrow \infty} g_n d\mu = \int_X \liminf_{n \rightarrow \infty} f_n d\mu. \quad (6.29)$$

It also follows that

$$g_n(x) \leq f_k(x) \quad \text{for any } k \geq n.$$

Hence,

$$\int_X g_n d\mu \leq \int_X f_k d\mu \quad \text{for any } k \geq n,$$

that is,

$$\int_X g_n d\mu \leq \inf_{k \geq n} \int_X f_k d\mu.$$

Taking the limit  $n \rightarrow \infty$  and applying the monotone convergence theorem, we get

$$\lim_{n \rightarrow \infty} \int_X g_n d\mu \leq \lim_{n \rightarrow \infty} \left[ \inf_{k \geq n} \int_X f_k d\mu \right] = \liminf_{n \rightarrow \infty} \int_X f_n d\mu. \quad (6.30)$$

From (6.29) and (6.30), we conclude that

$$\int_X \liminf_{n \rightarrow \infty} f_n d\mu = \int_X f d\mu \leq \liminf_{n \rightarrow \infty} \int_X f_n d\mu. \quad \clubsuit$$

### 6.3.4 Dominated Convergence Theorem (II)

Our next task is to prove the dominated convergence theorem.

*Proof* Observe that  $f_n$  and  $f$  are Lebesgue integrable on  $X$ . From hypothesis, it follows that  $f_n + g \geq 0$  and  $g - f_n \geq 0$  almost everywhere. Thus by Fatou's lemma, we have

$$\int_X \liminf_{n \rightarrow \infty} (f_n + g) d\mu \leq \liminf_{n \rightarrow \infty} \int_X (f_n + g) d\mu$$

or

$$\int_X \liminf_{n \rightarrow \infty} f_n d\mu \leq \liminf_{n \rightarrow \infty} \int_X f_n d\mu \quad (6.31)$$

by the linearity of the Lebesgue integral. It is also true that  $g - f_n \geq 0$  on  $X$ ; thus also by Fatou's lemma we have

$$\int_X \liminf_{n \rightarrow \infty} (g - f_n) d\mu \leq \liminf_{n \rightarrow \infty} \int_X (g - f_n) d\mu,$$

or equivalently,

$$-\int_X \liminf_{n \rightarrow \infty} f_n d\mu \leq \liminf_{n \rightarrow \infty} \left[ -\int_X f_n \right] d\mu.$$

The latter inequality can be rewritten as

$$\int_X \liminf_{n \rightarrow \infty} f_n d\mu \geq \limsup_{n \rightarrow \infty} \int_X f_n d\mu. \quad (6.32)$$

From (6.31) and (6.32) we set

$$\begin{aligned} \int_X \liminf_{n \rightarrow \infty} f_n d\mu &\leq \liminf_{n \rightarrow \infty} \int_X f_n d\mu \\ &\leq \limsup_{n \rightarrow \infty} \int_X f_n d\mu \leq \int_X \liminf_{n \rightarrow \infty} f_n d\mu, \end{aligned}$$

which clearly indicates that

$$\liminf_{n \rightarrow \infty} \int_X f_n d\mu = \limsup_{n \rightarrow \infty} \int_X f_n d\mu,$$

so that the limit  $\lim_{n \rightarrow \infty} \int_X f_n d\mu$  exists and is equal to  $\int_X \lim_{n \rightarrow \infty} f_n d\mu = \int_X f d\mu$ . This completes the proof of the theorem. ♣

### 6.3.5 Fubini Theorem

For a function of several variables, we may define the Lebesgue integral by exactly the same process as for a function of one variable. In cases of two variables, for instance, a rectangle  $S = [a, b] \times [c, d]$  takes on the role of intervals, and we need only to imitate the definitions and methods that we used for functions of a single variable. We can develop the theory for the entire plane  $\mathbf{R}^2$  analogously to that for the real axis  $\mathbf{R}$ . In fact, all the consequences in Sect. 6.2 for Lebesgue integrable functions on a closed interval  $[a, b]$  are easily carried over to the corresponding propositions for the double integral on the rectangle  $S$  without modifying the actual proofs in Sect. 6.2, except for replacing  $f(x)$  by  $f(x, y)$ .

However, an important new problem arises here. If  $f$  is integrable on the rectangle  $S = [a, b] \times [c, d]$ , we have to determine whether the value of the integral

$$\iint_S f(x, y) dx dy \quad (6.33)$$

is equal to that of the repeated integrals

$$\int_c^d \left[ \int_a^b f(x, y) dx \right] dy \quad \text{and} \quad \int_a^b \left[ \int_c^d f(x, y) dy \right] dx.$$

This is true for continuous functions on  $S$ . But it is far from obvious that the existence of the double integral (6.33) guarantees the existence of either repeated integral.

The following example may lead the reader to consider the point mentioned above.

*Examples* Assume the function

$$f(x, y) = \begin{cases} \frac{x^2 - y^2}{(x^2 + y^2)^2} & \text{for } (x, y) \neq (0, 0), \\ 0 & \text{for } (x, y) = (0, 0), \end{cases} \quad (6.34)$$

and compute the repeated integrals

$$I_{yx} = \int_0^1 dy \left[ \int_0^1 f(x, y) dx \right] \quad \text{and} \quad I_{xy} = \int_0^1 dx \left[ \int_0^1 f(x, y) dy \right].$$

Straightforward calculations yield

$$I_{xy} = \int_0^1 dx \int_0^1 \frac{\partial}{\partial y} \left( \frac{y}{x^2 + y^2} \right) dy = \int_0^1 \frac{dx}{x^2 + 1} dx = \frac{\pi}{4}$$

and

$$I_{yx} = \int_0^1 dy \int_0^1 \frac{\partial}{\partial x} \left( \frac{-x}{x^2 + y^2} \right) dx = \int_0^1 \frac{-dy}{y^2 + 1} dy = -\frac{\pi}{4}.$$

Hence, we conclude that

$$I_{xy} \neq I_{yx},$$

which indicates that the order of integrations with respect to  $x$  and  $y$  cannot be changed.

We now present the main theorem of this subsection.

♠ **Fubini theorem:**

Let the function  $f(x)$  be integrable on a rectangle  $S = [a, b] \times [c, d]$ . Then the following equalities hold:

$$\int \int_S f(x, y) dx dy = \int_c^d \left[ \int_a^b f(x, y) dx \right] dy = \int_a^b \left[ \int_c^d f(x, y) dy \right] dx.$$

According to Fubini's theorem, a double integral  $\int \int_S f(x, y) dx dy$  is computed by integrating first with respect to  $x$  and then with respect to  $y$ , or vice versa. We omit an exact proof of the Fubini theorem, since it requires rather lengthy arguments regarding the existence and the convergence of the double integrals. Instead, we present some applications of the theorem.

The following is an extension of the Fubini theorem:

♠ **Fubini–Hobson–Tonelli theorem:**

Let the function  $f(x)$  be defined on  $S = [a, b] \times [c, d]$ . Then, if either of the repeated integrals

$$\int_a^b \left[ \int_c^d |f(x, y)| dy \right] dx \quad \text{or} \quad \int_c^d \left[ \int_a^b |f(x, y)| dx \right] dy$$

exists,  $f$  is integrable on  $S$  and, hence,

$$\int \int_S f(x, y) dx dy = \int_a^b \left[ \int_c^d f(x, y) dy \right] dx = \int_c^d \left[ \int_a^b f(x, y) dx \right] dy.$$

Both the Fubini theorem and the Fubini–Hobson–Tonelli theorem for integrals on a rectangle  $S$  may be easily extended to integrals on all of  $\mathbf{R}^2$  or to the integrals on any measurable subsets of  $\mathbf{R}^2$ .

## Exercises

1. Suppose that the function

$$g_n(x) = -2k^2 x e^{-k^2 x^2} + 2(k+1)^2 x e^{-(k+1)^2 x^2}$$

is defined on  $[0, \infty)$ , and form the sum

$$f_n(x) = \sum_{k=1}^n g_k(x) = -2x e^{-x^2} + 2(n+1)^2 x e^{-(n+1)^2 x^2}.$$

Show that  $\int_0^\infty \lim_{n \rightarrow \infty} f_n(x) dx \neq \lim_{n \rightarrow \infty} \int_0^\infty f_n(x) dx$ .

**Solution:** We have

$$\int_0^\infty \lim_{n \rightarrow \infty} f_n(x) dx = \int_0^\infty (-2x e^{-x^2}) dx = [e^{-x^2}]_0^\infty = -1,$$

whereas

$$\lim_{n \rightarrow \infty} \int_0^\infty f_n(x) dx = [e^{-x^2} - e^{-(n+1)^2 x^2}]_0^\infty = 0.$$

Therefore, (6.23) is not valid. ♣

2. Given the function:

$$f_n(x) = \begin{cases} n \sin nx & \text{for } 0 \leq x \leq \pi/n, \\ 0 & \text{for } \pi/n \leq x \leq \pi, \end{cases}$$

show that

$$\lim_{n \rightarrow \infty} \int_0^1 f_n(x) dx \neq \int_0^1 \lim_{n \rightarrow \infty} f_n(x) dx.$$

**Solution:** We have  $\lim_{n \rightarrow \infty} f_n(x) = 0$  for every  $x$  in  $[0, \pi]$  and  $\lim_{n \rightarrow \infty} \int_0^\pi f_n(x) dx = 2$ . Hence, we obtain the desired result. ♣

3. Suppose that the nonnegative functions  $\{f_n(x) : n \in \mathbf{N}\}$  are each summable over a measurable set  $X$ , and  $f_n \leq f_{n+1}$  on  $X$ . Show that the limit function  $f = \lim_{n \rightarrow \infty} f_n$  is summable over  $X$  and that

$$\lim_{n \rightarrow \infty} \int_X f_n d\mu = \int_X f d\mu.$$

**Solution:** Let  $g_n = f_1 - f_n$ , so that  $0 = g_1 \leq g_2 \leq \cdots \leq f_1$ . Thus, the dominated convergence theorem ensures that  $\lim_{n \rightarrow \infty} g_n = f_1 - f$  is integrable, and we have  $\lim_{n \rightarrow \infty} \int_X (f_1 - f_n) d\mu = \int_X (f_1 - f) d\mu$ , which gives

$$\int_X f_1 d\mu - \lim_{n \rightarrow \infty} \int_X f_n d\mu = \int_X (f_1 - f) d\mu.$$

Further, as  $f$  is integrable since  $0 \leq f \leq f_1$ , we have

$$\int_X (f_1 - f) d\mu = \int_X f_1 d\mu - \int_X f d\mu,$$

so that

$$\int_X f_1 d\mu - \lim_{n \rightarrow \infty} \int_X f_n d\mu = \int_X f_1 d\mu - \int_X f d\mu,$$

which gives

$$\lim_{n \rightarrow \infty} \int_X f_n d\mu = \int_X f d\mu. \quad \clubsuit$$

4. Examine the applicability to integrals  $\int_0^\infty f_n(x) dx$  of dominated and monotone convergence theorems for the following: (i)  $f_n(x) = 2n^2 e^{-n^2 x^2}$ ; (ii)  $f_n(x) = n x e^{-n x^2}$

**Solution:**

(i) Setting  $y = nx$ , we have

$$\int_0^\infty f_n(x)dx = \int_0^\infty 2n^2 e^{-n^2 x^2} dx = \int_0^\infty 2n e^{-y^2} dy = n\sqrt{\pi},$$

where the last term diverges as  $n \rightarrow \infty$ . Hence, the  $\lim_{n \rightarrow \infty} \int_0^\infty f_n(x)dx$  does not exist. Next, we observe that for  $x \neq 0$ ,

$$\lim_{n \rightarrow \infty} f_n(x) = \lim_{n \rightarrow \infty} (2n^2 e^{-n^2 x^2}) = 0,$$

whereas for  $x = 0$ ,

$$\lim_{n \rightarrow \infty} f_n(0) = \lim_{n \rightarrow \infty} (2n^2) = \infty.$$

Thus, there is no limiting function  $f = \lim_{n \rightarrow \infty} f_n$  that satisfies the inequality  $f(x) \geq 2n^2 e^{-n^2 x^2}$  for all  $n$  in  $X$ , and we can conclude that neither the dominated nor the monotone convergence theorem is applicable.

(ii) It is found that

$$\int_0^\infty f_n(x)dx = \int_0^\infty nxe^{-nx^2} dx = \left[ -\frac{1}{2}e^{-nx^2} \right]_0^\infty = \frac{1}{2},$$

and that  $nxe^{-nx^2} \rightarrow 0$  pointwise as  $n \rightarrow \infty$ . Therefore, the limiting function  $f(x)$  satisfying the inequality  $f(x) \geq nxe^{-nx^2}$  does not exist. Hence, neither the dominated nor the monotone convergence theorem is applicable. ♣

**5.** Using Fubini's theorem, derive the formula

$$\int_0^1 \frac{x^b - x^a}{\log x} dx = \log \frac{1+b}{1+a} \quad \text{for } a, b > 0. \quad (6.35)$$

**Solution:** Note that the integral in the left-hand-side is beyond elementary calculus, so that it is impossible to achieve (6.35) by straightforward calculations. Instead, we observe that

$$\int_0^1 dx \int_a^b x^y dy = \int_0^1 \frac{x^b - x^a}{\log x} dx$$

and

$$\int_a^b dy \int_0^1 x^y dx = \int_a^b \frac{dy}{y+1} = \log \frac{1+b}{1+a}.$$

Thus, when we apply the Fubini theorem to the double integral

$$\int \int_{[0 \leq x \leq 1, a \leq y \leq b]} x^y dx dy,$$

we obtain the desired result (6.35). ♣

6. Show that the function  $f(x, y)$  given in (6.34) in Sect. 6.3.2 is not integrable on  $[0, 1] \times [0, 1]$ .

**Solution:** It follows that

$$\begin{aligned} \int \int_{0 \leq x, y \leq 1} \left| \frac{x^2 - y^2}{(x^2 + y^2)^2} \right| dx dy &= 2 \int \int_{0 \leq x \leq y \leq 1} \frac{y^2 - x^2}{(x^2 + y^2)^2} dx dy \\ &= 2 \int_0^1 dy \int_0^y \frac{y^2 - x^2}{(x^2 + y^2)^2} dx dy = \int_0^1 \frac{dy}{y} = \infty. \end{aligned}$$

This means that the existence and equality of two repeated integrals do *not* guarantee the existence of the double integral. ♣

## 6.4 The Lebesgue Spaces $L^p$

### 6.4.1 The Spaces of $L^p$

We close this chapter by demonstrating the relevance of the Lebesgue integral theory to the functional analysis that we discussed in Chap. 4. The Lebesgue theory on integration enables us to introduce certain spaces of functions that have properties that are of great importance in analysis as well as in mathematical physics, in particular, quantum mechanics. These are the so-called  $L^p$  spaces of complex-valued functions  $f$  such that  $|f|^p$  is integrable.

We have already dealt with the concept of Hilbert space. In fact,  $L^2$  for any measure  $\mu$  satisfies the conditions for a Hilbert space. We begin with a short review of the definition of  $L^p$  spaces in terms of measure, and follow this by examining how the spaces possess vector space properties owing to the use of the Lebesgue integral.

Let  $p$  be a positive real number and let  $X$  be a measurable set in  $\mathbf{R}$ . The  $L^p$  space is defined as follows:

#### ♠ Definition of $L^p$ space:

The  $L^p$  space is a set of complex-valued Lebesgue measurable functions  $f(x)$  on  $X$  that satisfy

$$\int_X |f|^p d\mu < \infty$$

for  $p \geq 1$ .

When the integral  $\int_X |f(x)|^p dx$  exists, we call it the  **$p$ -norm** of  $f$  and denote it by

$$\|f\|_p = \left( \int_X |f|^p d\mu \right)^{1/p}.$$

Clearly for  $p = 2$ , the present definition reduces to our earlier definition of  $L^2$ .

### 6.4.2 Hölder Inequality

The following two inequalities are fundamentals that demonstrate the relations between the norms of functions involved in  $L^p$ .

♠ **Hölder inequality:**

For any  $f, g \in L^p$  under the conditions

$$p, q > 1 \quad \text{and} \quad \frac{1}{p} + \frac{1}{q} = 1,$$

we have

$$fg \in L^1 \quad \text{and} \quad \|fg\|_1 \leq \|f\|_p \|g\|_q.$$

*Proof* We assume that neither  $f$  nor  $g$  is zero almost everywhere (otherwise, the result is trivial). To proceed with the proof, we first observe the inequality

$$a^{1/p} b^{1/q} \leq \frac{a}{p} + \frac{b}{q} \quad \text{for } a, b \geq 0, \quad (6.36)$$

which we justify by rewriting it as

$$t^{1/p} \leq \frac{t}{p} + \frac{1}{q},$$

where we set  $t = a/b$ . Then, we note that the function given by

$$f(t) = t^{1/p} - \frac{t}{p} - \frac{1}{q} \leq 0$$

has a maximum at  $t = 1$ , namely,

$$\max f(t) = f(1) = 1 - \frac{1}{p} - \frac{1}{q} = 0,$$

which results in the inequality (6.36), which we use to obtain

$$\frac{|f(x)g(x)|}{AB} \leq \frac{A^{-p}|f(x)|^p}{p} + \frac{B^{-q}|g(x)|^q}{q}, \quad (6.37)$$

where

$$A = \left[ \int_X |f|^p d\mu \right]^{1/p} \quad \text{and} \quad B = \left[ \int_X |g|^q d\mu \right]^{1/q}.$$

The right-hand side of (6.37) is integrable from the hypothesis that  $f, g \in L^p$ . Therefore, using (6.37) we obtain

$$\begin{aligned} \frac{1}{AB} \int_X |fg| d\mu &\leq \frac{A^{-p}}{p} \int_X |f|^p d\mu + \frac{B^{-q}}{q} \int_X |g|^q d\mu \\ &= \frac{1}{p} + \frac{1}{q} = 1. \end{aligned}$$

Consequently, we have

$$\int_X |fg| d\mu \leq AB,$$

which proves the inequality. ♣

### 6.4.3 Minkowski Inequality

The other inequality of interest is stated below.

#### ♠ Minkowski inequality:

If  $f, g \in L^p$  with  $p \geq 1$ , then

$$f + g \in L^p \quad \text{and} \quad \|f + g\|_p \leq \|f\|_p + \|g\|_p. \quad (6.38)$$

*Proof* For  $p = 1$ , the inequality is readily obtained by integrating the triangle inequality for real numbers. For  $p > 1$ , it follows that

$$\begin{aligned} \int_X |f + g|^p d\mu &= \int_X |f + g|^{p-1} |f| d\mu \\ &\quad + \int_X |f + g|^{p-1} |g| d\mu. \end{aligned}$$

Let  $q > 0$  be such that

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Applying the Hölder inequality to each of these last two integrals and noting that  $(p-1)q = p$ , gives us

$$\begin{aligned} \int_X |f(x) + g(x)|^p dx &\leq M \left[ \int_X |f + g|^{(p-1)q} d\mu \right]^{1/q} \\ &= M \left[ \int_X |f + g|^p d\mu \right]^{1/q}, \end{aligned} \quad (6.39)$$

where  $M$  denotes the right-hand side of the inequality (6.38) that we would like to prove. Now divide the extreme ends of the relation (6.39) by

$$\left[ \int_X |f + g|^p d\mu \right]^{1/q}$$

to obtain the desired result. ♣

**Remark.** It should be noted that neither the Hölder inequality nor the Minkowski inequality holds for  $0 < p < 1$  if  $\mu(X) > 0$ , which is why we restrict ourselves to  $p \geq 1$ .

#### 6.4.4 Completeness of $L^p$ Spaces

By virtue of the two inequalities discussed above, we can show the completeness properties of  $L^p$  spaces, which is crucially important for developing **Hilbert space theory** for Lebesgue measurable functions.

##### ♠ Completeness of $L^p$ spaces:

The space  $L^p$  is complete: i.e., for any  $f_n \in L^p$  satisfying

$$\lim_{n,m \rightarrow \infty} \|f_n - f_m\|_p = 0,$$

there exists  $f \in L^p$  such that

$$\lim_{n \rightarrow \infty} \|f_n - f\|_p = 0.$$

*Proof* Let  $\{f_n\}$  be a **Cauchy sequence** in  $L^p$ . Then, there is a natural number  $n_1$  such that for all  $n > n_1$ , we have

$$\|f_n - f_{n_1}\| < \frac{1}{2}.$$

By induction, after finding  $n_{k-1} > n_{k-2}$ , we find  $n_k > n_{k-1}$  such that for all  $n > n_k$  we have

$$\|f_n - f_{n_k}\| < \frac{1}{2^k}.$$

Then  $\{f_{n_k}\}$  is a subsequence of  $\{f_n\}$  that satisfies

$$\|f_{n_{k+1}} - f_{n_k}\| < \frac{1}{2^k}$$

or

$$\|f_{n_1}\| + \sum_{k=1}^{\infty} \|f_{n_{k+1}} - f_{n_k}\| = A < \infty.$$

Let

$$g_k = |f_{n_1}| + |f_{n_2} - f_{n_1}| + \cdots + |f_{n_{k+1}} - f_{n_k}|, \quad k = 1, 2, \dots.$$

Then, by the Minkowski inequality,

$$\begin{aligned} \int_X g_k^p(x) d\mu &= \int_X (|f_{n_1}| + |f_{n_2} - f_{n_1}| + \cdots + |f_{n_{k+1}} - f_{n_k}|)^p d\mu \\ &\leq \left( \|f_{n_1}\|_p + \sum_{k=1}^{\infty} \|f_{n_{k+1}} - f_{n_k}\| \right)^p \\ &\leq A^p < \infty. \end{aligned}$$

Let  $g = \lim g_k$ . Then  $g^p = \lim g_k^p$ . By the monotone convergence theorem given in Sect. 6.2.1, we have

$$\int_X g^p d\mu = \lim_{k \rightarrow \infty} \int_X g_k^p d\mu < \infty,$$

which shows that  $g$  is in  $L^p$ , and hence

$$\int_X \left( |f_{n_1}| + \sum_{k=1}^{\infty} |f_{n_{k+1}} - f_{n_k}| \right)^p dx < \infty,$$

implying that

$$|f_{n_1}| + \sum_{k=1}^{\infty} |f_{n_{k+1}} - f_{n_k}|$$

converges almost everywhere to a function  $f \in L^p$ .

It remains to prove that  $\|f_{n_k} - f\| \rightarrow 0$  as  $k \rightarrow \infty$ . We first note that

$$f(x) - f_{n_j}(x) = \sum_{k=j}^{\infty} [f_{n_{k+1}}(x) - f_{n_k}(x)].$$

It then follows that

$$\|f - f_{n_j}\| \leq \sum_{k=j}^{\infty} \|f_{n_{k+1}} - f_{n_k}\|_p < \sum_{k=j}^{\infty} \frac{1}{2^k} = \frac{1}{2^{j-1}}.$$

Therefore,  $\|f - f_{n_j}\|_p \rightarrow 0$  as  $j \rightarrow \infty$ . Now

$$\|f_n - f\|_p \leq \|f_n - f_{n_k}\|_p + \|f_{n_k} - f\|_p,$$

where  $\|f_n - f_{n_k}\|_p \rightarrow 0$  as  $n \rightarrow \infty$  and  $k \rightarrow \infty$  and thus  $\|f_n - f\|_p = 0$  as  $n \rightarrow \infty$ . This shows that the Cauchy sequence  $\{f_n\}$  converges to  $f$  in  $L^p$ . ♣

Before closing this chapter, we must emphasize that if we employ the Riemann integral to construct  $L^p$  spaces, the theorem mentioned above breaks down so that we can no longer expect completeness of the resulting function space. To illustrate this point, we temporarily define the ' $L^1$  space' by a set of Riemann integrable functions under the '1-norm':

$$\|f\|_1^{[R]} \equiv \int_0^1 |f(x)| dx < \infty.$$

We then consider a function

$$f_n(x) = \begin{cases} 1 & \text{for } x \in \{a_n\}, \\ 0 & \text{otherwise,} \end{cases}$$

where  $\{a_n\}$  ( $n = 1, 2, \dots$ ) is an infinite sequence of all rational numbers in  $[0, 1]$ . It readily follows that a function  $f_n(x) - f_\mu(x) \in L^1$  is Riemann integrable and reads

$$\|f_n - f_\mu\|_1^{[R]} = \int_0^1 |f_n(x) - f_\mu(x)| dx = 0.$$

Nevertheless,  $f_n(x)$  converges to Dirichlet's function  $\chi(x)$ , which is not Riemann integrable as noted earlier. As it is impossible to examine the quantity

$$\|f_n - \chi\|_1^{[R]},$$

using Riemann integrals, we cannot establish the complete function space based on that method.

## 6.5 Applications in Physics and Engineering

### 6.5.1 Practical Significance of Lebesgue Integrals

From a practical viewpoint, what makes Lebesgue integrals so important is the fact that they allow us to interchange the order of integration and other limiting procedures under very weak conditions, which is not possible in the case of Riemann integrals. In fact, in the case of Riemann integrals, the identities

$$\lim_{n \rightarrow \infty} \int_{-\infty}^{\infty} f_n(x) dx = \int_{-\infty}^{\infty} \lim_{n \rightarrow \infty} f_n(x) dx$$

and

$$\sum_{n=1}^{\infty} \int_{-\infty}^{\infty} f_n(x) dx = \int_{-\infty}^{\infty} \sum_{n=1}^{\infty} f_n(x) dx$$

are valid only if the integrands on the right-hand side, i.e.,  $\lim f_n$  and  $\sum f_n$ , are uniformly convergent. Such a restriction can be removed by using a Lebesgue

integral since with the latter, only pointwise convergence of the integrand is needed. We saw in Sect. 6.3 that the **Lebesgue convergence theorem** and **Fubini's theorem** markedly weaken the conditions necessary for the validity of an interchange of the order of integration. As a result, we need not monitor the order of the limiting procedure, which is very useful in the practical calculations encountered in physics and engineering.

### 6.5.2 Contraction Mapping

Another important consequence of Lebesgue integral theory is the **completeness** of the **function space**  $L^p$  spanned by Lebesgue integrable functions.  $L^p$  spaces have a wide range of applications in physics, statistics, engineering, and other disciplines. For instance, they serve as a basis in the development of a rigorous theory of Fourier transformation, in which the mappings between two different  $L^p$  spaces are considered. Moreover, the theory of quantum mechanics is established on the basis of the  $L^2$  space, a specific class of  $L^p$  spaces with  $p = 2$ . In both applications, the completeness property of the  $L^p$  space plays a crucial role in making the theory self-contained. In order for the reader to learn more about this issue, we present the **contraction mapping theorem** (or **Banach's fixed point theorem**) below. This theorem proves the existence of a unique solution to a certain kind of equation associated with Lebesgue integrable functions, which makes the theory based on  $L^p$  spaces self-contained.

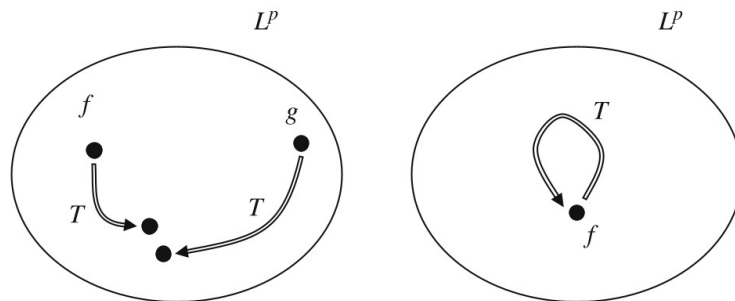
A preliminary terminology is defined below.

♠ **Contraction mapping:**

A contraction mapping  $T$  is a mapping from  $L^p$  onto  $L^p$  that satisfies the relation

$$\|T(f) - T(g)\| \leq c\|f - g\| \quad (0 \leq c < 1) \quad (6.40)$$

for any  $f, g \in L^p$  (see Fig. 6.6).



**Fig. 6.6.** Sketch of a contraction mapping  $T$  acting on  $f, g \in L^p$

**Remark.** If  $T$  is regarded as a differential operator acting on a Lebesgue integrable function  $f$ , then we can say that ‘a contraction mapping is a mapping that satisfies the **Lipschitz condition**’ (see Sect. 15.2.3).

We should keep in mind that the norm  $\|\cdots\|$  used in (6.40) is in terms of  $L^p$  spaces, so that  $\|f - g\| = 0$  means  $f = g$  **almost everywhere**. In plain words, a contraction mapping reduces the distance between two elements in the  $L^p$  space.

We are now ready to move on to the main theorem.

♠ **Contraction mapping theorem:**

Let  $T$  be a contraction mapping and  $I$  be an identity mapping. Then the equation

$$(T - I)f = 0 \quad (6.41)$$

has one and only one solution  $f$  that belongs to  $L^p$ .

**Remark.** The solution  $f$  of the equation (6.41) is called a **fixed point** in  $L^p$ .

The contraction mapping theorem guarantees the existence and uniqueness of fixed points of certain self-mappings and provides a constructive method for finding those fixed points. It should be emphasized that the theorem allows us to prove the existence (and uniqueness) of solutions of **ordinary differential equations** with respect to Lebesgue integrable functions, as intuitively understood if  $T$  is set to be a differential operator.

**Proof (of the contraction mapping theorem):** For arbitrary  $f_0 \in L^p$ , we introduce a sequence of functions  $\{f_n\}$  defined by

$$f_1 = T(f_0), f_2 = T(f_1), \cdots, f_n = T(f_{n-1}), \cdots$$

We shall see below that the sequence  $\{f_n\}$  is a **Cauchy sequence** and thus has a limit  $f \equiv \lim_{n \rightarrow \infty} f_n$ . It follows from the definition of  $T$  that

$$\begin{aligned} \|f_n - f_{n+j}\| &= \|T(f_{n-1}) - T(f_{n-1+j})\| \\ &\leq c \|f_{n-1} - f_{n-1+j}\| \\ &\leq \cdots \leq c^n \|f_0 - f_j\| \end{aligned} \quad (6.42)$$

and

$$\begin{aligned} \|f_0 - f_j\| &\leq \|f_0 - f_1\| + \cdots + \|f_{j-1} - f_j\| \\ &\leq (1 + c + \cdots + c^{j+1}) \|f_0 - f_1\| \\ &\leq (1 - c)^{-1} \|f_0 - f_1\|, \end{aligned} \quad (6.43)$$

where we used the **Minkowski inequality** (6.38) with respect to the  $p$ -norm designated by  $\|\cdots\|$ . From (6.42) and (6.43), we set

$$\|f_n - f_{n+j}\| \leq \frac{c^n}{1-c} \|f_0 - f_1\| \rightarrow 0 \quad (n \rightarrow \infty).$$

This indicates that  $\{f_n\}$  is a Cauchy sequence and thus converges to a limit (denoted by  $f$ ) regardless of the choice of  $f_0$ . Furthermore, the limit  $f$  always belongs to  $L^p$  since the space  $L^p$  is complete. Hence, the converging behavior of  $\{f_n\}$  to  $f$  can be expressed by using the concept of the norm of  $L^p$  as

$$\lim_{n \rightarrow \infty} \|f_n - f\| = 0. \quad (6.44)$$

We then obtain

$$\begin{aligned} \|T(f) - f\| &\leq \|T(f) - f_n\| + \|f_n - f\| \\ &= \|T(f) - T(f_{n-1})\| + \|f_n - f\| \\ &\leq \|f - f_{n-1}\| + \|f_n - f\| \rightarrow 0 \quad (n \rightarrow \infty), \end{aligned} \quad (6.45)$$

which means that  $T(f) = f$  almost everywhere. Consequently, equation (6.41) has at least one solution that is a limit  $f$  of the sequence  $\{f_n\}$  that we introduced.

The uniqueness of the solution  $f$  is readily understood. Suppose  $g \in L^p$  such that  $T(g) = g$ . We then have

$$\|g - f\| = \|T(g) - T(f)\| \leq c\|g - f\|.$$

This means that  $\|g - f\| = 0$  since  $0 \leq c < 1$ , so we have  $g = f$  almost everywhere. ♣

**Remark.** Note that it is our use of the Lebesgue integral (instead of the Riemann integral) that guarantees the validity of the contraction mapping theorem. In fact, if we restrict ourselves to the Riemann integral, the limit  $f$  of the sequence  $\{f_n\}$  may not belong to  $L^p$ , and we can no longer obtain the result (6.45).

### 6.5.3 Preliminaries for the Central Limit Theorem

The effectiveness of Lebesgue integrals is also observed in probability theory, particularly in the derivation of the **central limit theorem**, which plays a fundamental role in statistical mechanics and in the statistical analysis of experimental data. Later, we shall see that employing Lebesgue integrals is necessary for proving the central limit theorem, where the **Lebesgue convergence theorem** and **Fubini's theorem** are used time and again.

In order to prove the central limit theorem, we introduce a **random variable**  $x$  (see Sect. 6.1.3); for instance,  $x$  may be the number of spots we get when shooting a pair of dice or a real number that we randomly pick from an interval on the real axis. Suppose that  $x$  lies in a set  $X$  on the real axis. (Here,  $X$  may be a continuous interval, a set of discrete points, or a union of the two.) In modern probability theory, measures characterizing the statistical properties of the system considered are defined in terms of the Lebesgue integral. For instance, the **probability** (or **distribution**) that  $x$  is found in subset  $X_0 \subset X$  is given by

$$P(x : x \in X_0) = \int_{X_0} p d\mu, \quad (6.46)$$

where  $\mu$  is the Lebesgue measure of  $X_0$  and  $p$  is the **probability density** associated with  $x$ . In general,  $p$  is assumed to satisfy the normalization condition  $\int_X p d\mu = 1$ . We can state that the random variables  $x$  and  $y$  are **independent** if

$$P(x, y) = P(x)P(y).$$

Moreover, the variables  $x$  and  $y$  are said to be **identically distributed** if

$$P(x) = P(y).$$

We also define the **expected** (or **mean**) **value** of  $x$  and the variance of  $x$  by the integrals

$$E\{x\} = \int_X x p d\mu \quad \text{and} \quad V\{x\} = \int_X (x - E\{x\})^2 p d\mu,$$

respectively, where  $\mu$  is the Lebesgue measure of  $X$ . In particular, the expected value of an imaginary exponent  $e^{izx}$ , where  $z$  is real, is known as the **characteristic function**.

♠ **Characteristic function:**

The characteristic function  $\varphi_x(z)$  of a random variable  $x$  is defined by

$$\varphi_x(z) = E\{e^{izx}\}.$$

It can be shown that

$$E\{e^{iz(x+y)}\} = \varphi_x(z)\varphi_y(z)$$

if and only if the random variables  $x$  and  $y$  are independent. Furthermore, we obtain

$$\varphi_x(z) = \varphi_y(z)$$

if and only if the variables  $x$  and  $y$  are identically distributed. The latter condition is known as the **uniqueness theorem** for characteristic functions, and the proof, which involves **Fubini's theorem**, can be found in advanced texts on probability theory.

### 6.5.4 Central Limit Theorem

We are now ready to state the key theorem.

♠ **Central limit theorem:**

Assume a series of random variables  $\{x_n\}$  in which the  $x_n$  are independent and identically distributed. For arbitrary  $a$  and  $b$  ( $b > a$ ), we have

$$\lim_{n \rightarrow \infty} P \left( a \leq \frac{\sum_{j=1}^n x_j - nm}{\sigma \sqrt{n}} \leq b \right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\xi^2/2} d\xi, \quad (6.47)$$

where  $m = E\{x_n\}$  and  $\sigma^2 = V\{x_n\}$ .

Briefly, the theorem states that the probability that the average of  $n$  random variables equals  $\alpha$  is proportional to  $e^{-\alpha^2/2}$ . (Note that  $\alpha$  is the *average* of  $n$  variables and not a variable itself.) A random variable with the probability density  $e^{-\xi^2/2}$  is said to be **normally distributed**.

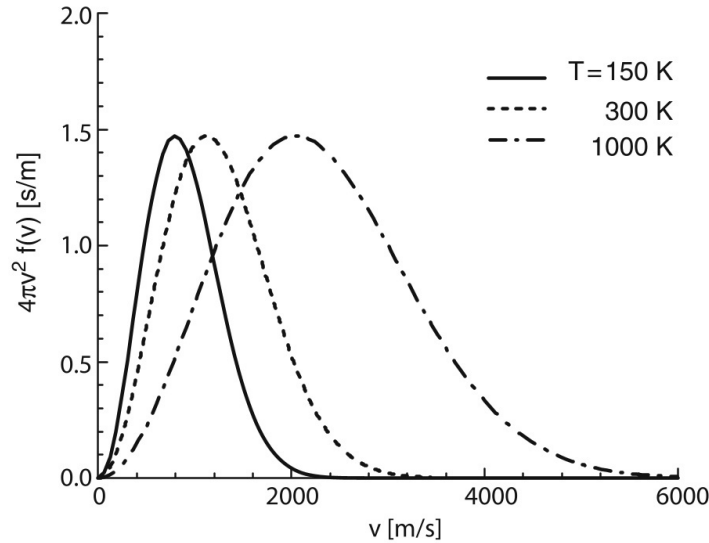
**Remark.** The central limit theorem is very effective in describing various stochastic phenomena in nature since it can be applied regardless of the distribution of the  $n$  random variables; i.e., almost all classes of random variables obey the theorem as long as they are independent and identically distributed.

An illustrative example of the central limit theorem in physics is the **Maxwell – Boltzmann distribution** of an ideal gas. For a given temperature  $T$ , the distribution  $f(v)$  of the velocity of gas molecules  $v = |\mathbf{v}|$  is known to satisfy the equation

$$f(v) = \left( \frac{m}{2\pi k_B T} \right)^{3/2} \exp \left( -\frac{mv^2}{2k_B T} \right), \quad (6.48)$$

where  $m$  is the mass of a gas molecule and  $k_B$  is the Boltzmann constant. Here, the velocity  $v(t_i)$  as a function of discrete time  $t_i$  ( $i = 1, 2, \dots, n$ ) serves as  $n$  random variables. In general, in an equilibrium state,  $v(t_i)$  for different  $t_i$  is independent and identically distributed and, thus, if  $n$  is sufficiently large, the time average of  $v(t_i)$  obeys the normal distribution described by (6.48). Figure 6.7 shows the distribution of the squared velocity of gas molecules, which is determined from the formula  $4\pi v^2 f(v^2)$ , for various values of  $T$ ; we set  $k_B = 1.38 \times 10^{-23} \text{ kg} \cdot \text{m}^2/\text{s}^2 \cdot \text{K}$  and  $m = 6.6 \times 10^{-27} \text{ kg}$  by considering  $^4\text{He}$  molecules. We observe that the mean value of  $v^2$  shifts to the right with an increase in the temperature, which can be intuitively understood to be due to the acceleration of the molecules at high temperatures.

It is important to emphasize that the central limit theorem holds good for any kind of distribution of the  $n$  variables  $\{x_i\}$  as long as they are independent and identically distributed. For example, let us consider  $n$  variables that obey



**Fig. 6.7.** Distribution of square velocity  $v^2$  of  $^4\text{He}$  molecules

the distribution  $P(x)$  shown in Fig. 6.8. The average of these variables shows the distribution depicted in Fig. 6.8, all of which converge to the normal distribution as  $n$  increases. The fact that the distribution of  $\{x_i\}$  can be disregarded is the reason the normal distribution is so universally observed in a wide variety of stochastic phenomena.

### 6.5.5 Proof of the Central Limit Theorem

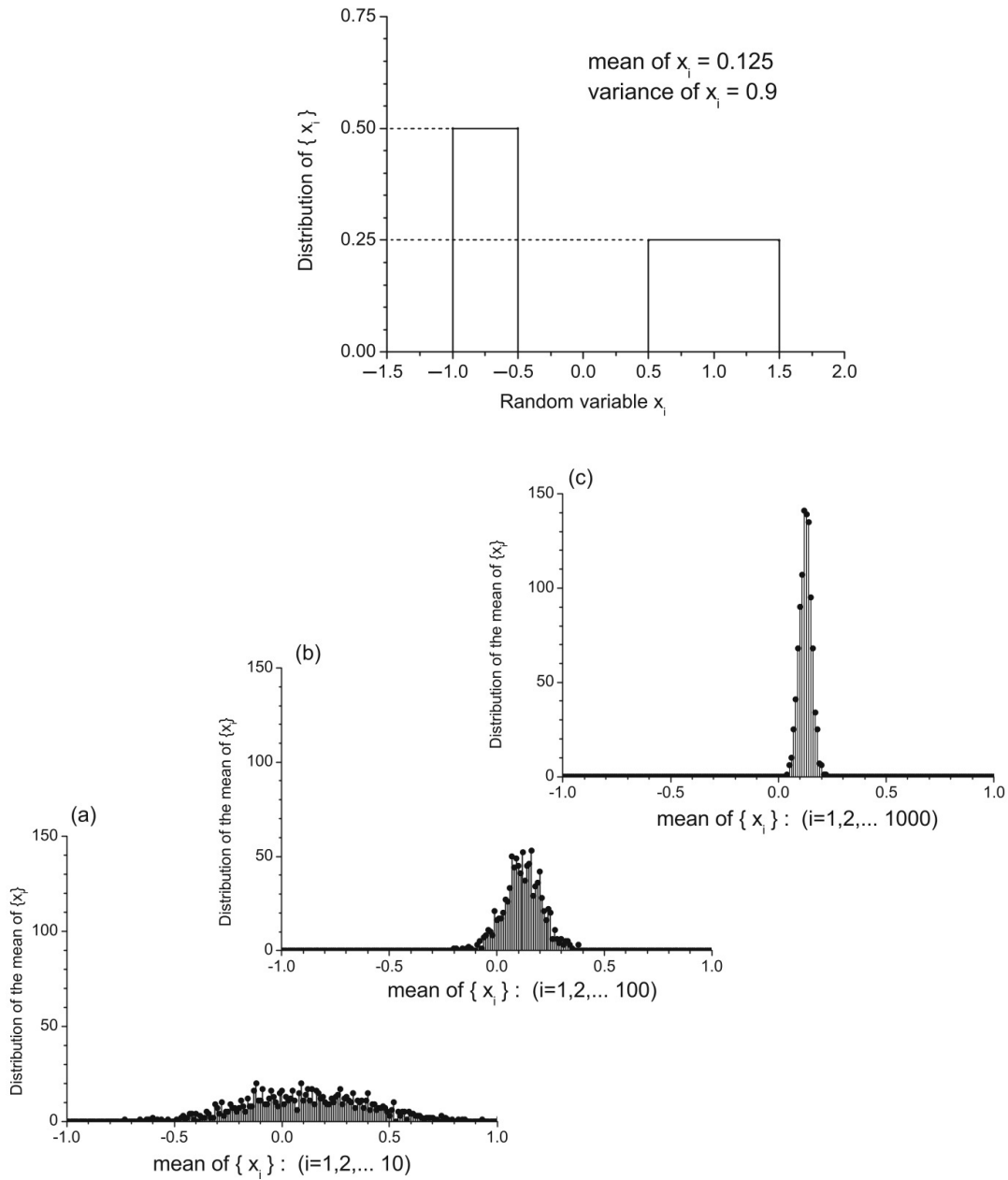
As some further points have to be discussed in order to prove the central limit theorem, we present below only an outline and not a rigorous proof. Let us emphasize that the use of Lebesgue integrals is necessary for proving the central limit theorem, and the **Lebesgue convergence theorem** and **Fubini's theorem** are used time and again.

*Proof* We have only to consider the case of  $m = 0$  and  $\sigma = 1$ ; otherwise, the new variable  $\tilde{x}_n \equiv (x_n - m)/\sigma$  is introduced to yield  $E\{\tilde{x}_n\} = 0$  and  $V\{\tilde{x}_n\} = 1$ . The characteristic function  $\varphi_{y_n}(z)$  for the variable

$$y_n = \frac{\sum_{j=1}^n x_j}{\sqrt{n}}$$

is given by

$$\varphi_{y_n}(z) = E \left\{ \exp \left( \frac{iz}{\sqrt{n}} \sum_{j=1}^n x_j \right) \right\} = \prod_{j=1}^n \varphi_{x_j} \left( \frac{z}{\sqrt{n}} \right),$$



**Fig. 6.8.** *Top:* Distributions of a random variable  $x$ . *Bottom:* (a)–(c) Distributions of the average value  $\alpha$  of  $n$  random variables  $x_1, x_2, \dots, x_n$  with  $n = 10$  for (a),  $n = 100$  for (b), and  $n = 1000$  for (c). For each, 1000  $\alpha$ 's are sampled to create the distribution. With increasing  $n$ , the distribution of  $\alpha$  converges to the normal distribution around the center of 0.125 as expected

in which the condition that all  $x_n$  are independent allows us to obtain the last expression. Furthermore, since all  $x_n$  are identically distributed, we have

$$\prod_{j=1}^n \varphi_{x_j} \left( \frac{z}{\sqrt{n}} \right) = \left[ \varphi_{x_1} \left( \frac{z}{\sqrt{n}} \right) \right]^n,$$

which is ensured by the **uniqueness theorem** discussed in Sect. 6.5.5.

We want the limit of  $\varphi_{y_n}(z)$  at  $n \rightarrow \infty$ , so we use the formula (see the lemma below)

$$\lim_{n \rightarrow \infty} \varphi_{y_n}(z) = e^{-z^2/2}, \quad (6.49)$$

in which the right-hand side is the characteristic function of a normal distribution. The result of (6.49) together with the **continuity theorem** (see below) states that

$$\lim_{n \rightarrow \infty} P(a \leq y_n \leq b) = P(a \leq y \leq b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-y^2/2} dy. \quad \clubsuit \quad (6.50)$$

The following theorem forms the basis for the proof of the central limit theorem.

**♠ Continuity theorem:**

Let  $x$  and  $x_n$  be random variables such that

$$\lim_{n \rightarrow \infty} \varphi_{x_n}(z) = \varphi_x(z).$$

We then obtain

$$\lim_{n \rightarrow \infty} P(a \leq x_n \leq b) = P(a \leq x \leq b)$$

for arbitrary  $a, b (b > a)$  satisfying  $P(x = a) = P(x = b) = 0$ .

This theorem states that the convergence of characteristic functions implies the convergence of the corresponding distribution functions. Since the proof requires the use of Fubini's theorem as well as the Lebesgue convergence theorem and is quite complicated, we do not present it.

**♠ Lemma:**

If  $E\{x\} = 0$  and  $E\{x^2\} = 1$  for a random variable  $x$ , then the characteristic function  $\varphi_x(z)$  satisfies the relation

$$\lim_{n \rightarrow \infty} \left[ \varphi_x \left( \frac{z}{\sqrt{n}} \right) \right]^n = \exp \left( -\frac{z^2}{2} \right). \quad (6.51)$$

*Proof* The assumption that  $E\{x^2\} = 1 < \infty$  implies that  $\varphi_x(z)$  is twice differentiable. In fact, we obtain

$$\begin{aligned}\varphi_x(z) &= E\{e^{izx}\}, \\ \varphi_x'(z) &= E\left\{\frac{d}{dz}e^{izx}\right\} = E\{ixe^{izx}\}, \\ \varphi_x''(z) &= E\left\{\frac{d^2}{dz^2}e^{izx}\right\} = E\{-x^2e^{izx}\},\end{aligned}$$

where the **Lebesgue convergence theorem** was used to interchange the order of differentiation  $d/dz$  and integration  $\int dx$  associated with calculation of  $E\{\cdots\}$ . The twice differentiability of  $\varphi(z)$  allows us to expand it around  $z = 0$  as

$$\varphi_x\left(\frac{z}{\sqrt{n}}\right) = \varphi_x(0) + \frac{z}{\sqrt{n}}\varphi_x'(0) + \frac{z^2}{2n}\varphi_x''(\eta),$$

where  $\eta$  is small enough to be  $|\eta| \leq |z|/\sqrt{n}$ . Since  $\varphi_x(0) = 1$  and  $\varphi_x'(0) = E\{ix\} = 0$ , we have

$$\begin{aligned}\log\left[\varphi\left(\frac{z}{\sqrt{n}}\right)\right]^n &= n\log\varphi\left(\frac{z}{\sqrt{n}}\right) \\ &= n\log\left(1 + \frac{z^2}{2n}\varphi''(\eta)\right) \\ &= \frac{z^2}{2}\varphi''(\eta) - \frac{z^4}{8n}\varphi''(\eta)^2 + \cdots \quad (n \gg 1),\end{aligned}$$

where we used the inequality  $|\varphi''(\eta)| \leq 1$  to expand the logarithmic term for  $n \gg 1$ . As a result, we set

$$\lim_{n \rightarrow \infty} \log\left[\varphi\left(\frac{z}{\sqrt{n}}\right)\right]^n = \frac{z^2}{2}\varphi''(0) = -\frac{z^2}{2},$$

which is equivalent to (6.51). ♣.



## Part III

---

### Complex Analysis



## Complex Functions

**Abstract** Differentiation and integration of complex functions are significantly different from those of real functions. In this chapter, we show that two very important theorems—the Cauchy theorem (Sect. 7.2.2) and the Taylor series expansion (Sect. 7.4.3)—result in a broad range of mathematical consequences that are highly relevant and useful in mathematical physics. However, before moving on to the principal discussion, we deal with the underlying concepts of analytic functions (Sect. 7.1.2) and the geometric meaning of analyticity (Sect. 7.1.5).

### 7.1 Analytic Functions

#### 7.1.1 Continuity and Differentiability

This chapter describes the theory of functions of a complex variable. Let  $\mathbf{C}$  denote the set of all elements  $z$  of the form

$$z = x + iy,$$

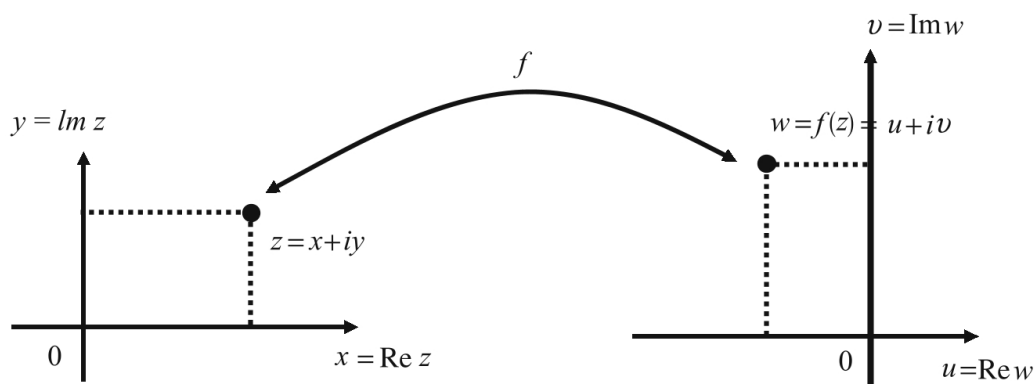
where  $x, y \in \mathbf{R}$  and  $i$  is a familiar symbol defined by  $i^2 = -1$ . Let  $D$  be a domain in  $\mathbf{C}$ . Then, a **complex function** defined by

$$f : D \rightarrow \mathbf{C}$$

is a rule that assigns a complex-valued function  $f(z)$  to each  $z \in D$ . This  $f(z)$  is equivalent to an ordered pair of real-valued functions  $u(z)$  and  $v(z)$ . Thus,  $f(z)$  can be written in the form

$$w = f(z) = u(z) + iv(z).$$

The real-valued functions  $u(z)$  and  $v(z)$  are called the **real** and **imaginary parts** (or **components**) of  $f(z)$  (see Fig. 7.1). We may write  $u = \operatorname{Re} f$  and  $v = \operatorname{Im} f$ .



**Fig. 7.1.** A complex function  $w = f(z)$  that assigns a point on the  $w$ -plane to each point on the  $z$ -plane

Once we introduce complex functions, the concepts of differentiation and integration encountered in ordinary real calculus acquire new depth and significance. When  $f(z)$  has its derivative in  $D$ , it is referred to as an **analytic function** in  $D$ . (More precise definitions of analytic functions are given in Sect. 7.1.3.) We shall see that the conditions for a complex-valued function  $f(z)$  to be differentiable with respect to a complex variable  $z$  is much stronger than that for a real-valued function  $f(x)$  with respect to a real variable  $x$ . This restriction forces a great deal of the structure of  $f(z)$ .

An exact definition of an analytic function is obtained by considering its derivative with respect to a complex variable  $z$ . Therefore, our first task is to determine the necessary and sufficient conditions for a complex function  $f(z)$  to have a derivative with respect to  $z$ . Before stating what is meant by the derivative  $f'(z)$ , we begin with the definition of continuity for  $f(z)$ .

♠ **Continuity of complex functions:**

Let  $f : D \rightarrow \mathcal{C}$  be a complex function and  $z_0$  a point in  $D$ . Then, a function  $w = f(z) \in \mathcal{C}$  is continuous at the point  $z_0$  if

$$\lim_{z \rightarrow z_0} f(z) = f(z_0). \quad (7.1)$$

In the limit of (7.1), the complex variable  $z$  may approach  $z_0$  from any direction in  $D$  (see Fig. 7.2). Hence, if we say the limit (7.1) exists, it means that a unique quantity  $f(z_0)$  must result from the limiting process regardless of how the limit  $z \rightarrow z_0$  is taken.

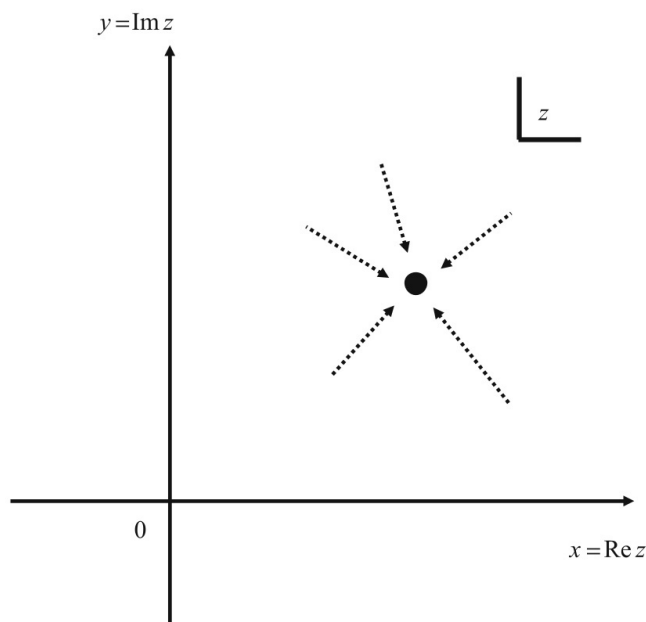
A similar feature is found in the definition of the derivative of  $f(z)$ .

♠ **Derivatives of complex functions:**

A complex function  $f(z)$  is said to be **differentiable** at the point  $z_0$  if and only if the limit

$$\lim_{z \rightarrow z_0} \frac{f(z) - f(z_0)}{z - z_0}, \quad z \in D \quad (7.2)$$

exists and is uniquely determined regardless of the manner in which  $z$  approaches  $z_0$ . When the limit exists, we denote it by  $f'(z_0)$ , the **derivative** of  $f(z)$  at  $z_0$ .



**Fig. 7.2.** Approaching direction of  $z$  to  $z_0$

The definition(7.2) requires that the ratio  $[f(z_0 + \Delta z) - f(z_0)]/\Delta z$  always tend to a unique limiting value, no matter the path along which  $z$  approaches  $z_0$ . This is an extremely strict condition; in fact, a number of theorems in the theory of analytic functions are derived from this requirement.

Keep in mind that a function  $f(z)$  may be differentiable only at a point, or on a curve, or through a region. An example for a differentiable function at single point is presented in Example 3 in Sect. 7.1.2.

### 7.1.2 Definition of an Analytic Function

Among many differentiable functions, some specific kinds of functions form the class of analytic functions as stated below.

#### ♠ Analytic functions:

A function  $f(z)$  is said to be **analytic** at the point  $z = z_0$  if and only if it is differentiable throughout a neighborhood of  $z = z_0$ .

**Remark.** There are some synonyms for the term analytic: **holomorphic**, **regular**, and **regular analytic**.

We offer some comments on the distinction between **differentiability** and **analyticity**. As noted above, the conditions for  $f(x)$  to be analytic are more stringent than those for it to be differentiable; in fact, a function  $f(z)$  is said to be analytic at a point  $z_0$  if it has a derivative at  $z_0$  and *at all points in some neighborhood*  $z_0$ . In this context, if we say that a function is analytic on a curve, we mean that it has a derivative at all points on a two-dimensional narrow strip containing the curve. If a function is differentiable only at a point or only along a curved line, then it is not analytic so that we say it is **singular** there. A typical example of  $f(x)$  that is differentiable only at a point is demonstrated in Example 3 below.

**Examples 1.** The function  $f(z) = z^n$  is differentiable and analytic everywhere. In fact, the limit

$$\begin{aligned} & \lim_{\Delta z \rightarrow 0} \frac{(z_0 + \Delta z)^n - z_0^n}{\Delta z} \\ &= \lim_{\Delta z \rightarrow 0} \left[ nz_0^{n-1} + \frac{n(n-1)}{2} z_0^{n-2} \Delta z + \cdots + (\Delta z)^{n-1} \right] = nz_0^{n-1} \end{aligned}$$

exists for arbitrary  $z_0$ , and is clearly independent of the path along which  $\Delta z \rightarrow 0$ . This means that any polynomial in  $z$  is differentiable and analytic everywhere.

**2.** The function  $f(z) = z^*$  is neither differentiable nor analytic anywhere, since the limit yields

$$\lim_{\Delta z \rightarrow 0} \frac{(z_0 + \Delta z)^* - z_0^*}{\Delta z} = \lim_{\Delta z \rightarrow 0} \frac{\Delta z^*}{\Delta z}. \quad (7.3)$$

If  $\Delta z \rightarrow 0$  parallel to the real axis, then  $\Delta z = \Delta z^* = \Delta x$  so that the limit equals 1. However, if  $\Delta z \rightarrow 0$  parallel to the imaginary axis, then  $\Delta z = i\Delta y = -i\Delta z^*$  so that the limit equals  $-1$ . Therefore, the quantity (7.3) depends on the path  $\Delta z \rightarrow 0$ , which means that it is neither differentiable nor analytic anywhere.

**3.** The function  $f(z) = |z|$  is differentiable only at the origin. In fact,

$$\begin{aligned} f(z_0 + \Delta z) - f(z_0) &= (z_0 + \Delta z)(z_0^* + \Delta z^*) - z_0 z_0^* \\ &= z_0 \Delta z^* + z_0^* \Delta z - \Delta z \Delta z^*, \end{aligned}$$

which yields

$$\begin{aligned} \lim_{\Delta z \rightarrow 0} \frac{f(z_0 + \Delta z) - f(z_0)}{\Delta z} &= \lim_{\Delta z \rightarrow 0} \frac{z_0 \Delta z^* + z_0^* \Delta z - \Delta z \Delta z^*}{\Delta z} \\ &= cz_0 + z_0^*, \end{aligned}$$

where  $c = \lim_{\Delta z \rightarrow 0} (\Delta z^* / \Delta z)$  is a complex-valued constant that depends on the path of  $\Delta z \rightarrow 0$ . Hence, the limit noted above is uniquely determined only when  $z_0 = 0$ , which means that the function  $f(z) = |z|$  is differentiable only at a point  $z = 0$ .

### 7.1.3 Cauchy–Riemann Equations

Let  $f : D \rightarrow \mathbf{C}$  with  $f(z) = u(z) + iv(z)$  as usual. We give the necessary and sufficient conditions for a function  $f(z) = u(x, y) + iv(x, y)$  to be differentiable at a point  $z_0 \in D$ . Let us assume that  $f(z)$  is differentiable at  $z_0 \in D$ . Then we have

$$f'(z_0) = \lim_{\Delta z \rightarrow 0} \frac{\Delta f}{\Delta z} = \lim_{\Delta z \rightarrow 0} \left( \frac{\Delta u}{\Delta z} + i \frac{\Delta v}{\Delta z} \right).$$

Since  $f'(z_0)$  exists, it is independent of the path  $\Delta z \rightarrow 0$ ; i.e., it is independent of the ratio  $\Delta y / \Delta x$ . If the limit is taken parallel to the real axis,  $\Delta y = 0$  and  $\Delta z = \Delta x$ , we have

$$f'(z_0) = \lim_{\Delta x \rightarrow 0} \left( \frac{\Delta u}{\Delta x} + i \frac{\Delta v}{\Delta x} \right) = \frac{\partial u}{\partial x} + i \frac{\partial v}{\partial x}.$$

On the other hand, if the limit approaches the point  $z_0$  along the line parallel to the imaginary axis,  $\Delta x = 0$  and  $\Delta z = i\Delta y$ , then

$$f'(z_0) = \lim_{\Delta y \rightarrow 0} \left( \frac{\Delta v}{\Delta y} - i \frac{\Delta u}{\Delta y} \right) = \frac{\partial v}{\partial y} - i \frac{\partial u}{\partial y}.$$

From the initial assumption, these two limits must be equal, so equating real and imaginary parts gives us

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \quad \text{and} \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}. \quad (7.4)$$

Equations (7.4) are known as the **Cauchy–Riemann relations** (abbreviated by CR relations), and they are a necessary condition for differentiability.

However, alone they are not sufficient, as they provide only necessary condition. This is because they were determined from special cases of the requirement of differentiability as demonstrated above. In fact, the sufficient conditions for the differentiability of  $f(z)$  at  $z_0$  consist of the following two statements:

♠ **Theorem:**

A function  $f(z)$  is differentiable at  $z_0 \in D$  if and only if

- (i) the first-order partial derivatives of  $u(x, y)$  and  $v(x, y)$  exist and are continuous at  $z_0$ , and
- (ii) those derivatives at  $z_0$  satisfy the CR equations.

*Proof* We prove that conditions **(i)** and **(ii)** imply the differentiability of  $f(z)$  at  $z_0 \in D$ . (The converse was proven implicitly in the beginning of this subsection.) From hypothesis **(i)**, the functions  $u$ ,  $\partial u/\partial x$ , and  $\partial u/\partial y$  are all continuous at the point  $z_0 = x_0 + iy_0$ , so we have

$$\begin{aligned}\Delta u &= u(x_0 + \Delta x, y_0 + \Delta y) - u(x_0, y_0) \\ &= \frac{\partial u}{\partial x} \Delta x + \frac{\partial u}{\partial y} \Delta y + \varepsilon_1 \Delta x + \varepsilon_2 \Delta y,\end{aligned}\tag{7.5}$$

in approximation of the order  $\Delta x$  and  $\Delta y$ . In (7.5), the partial derivatives are equated at the point  $(x_0, y_0)$ , and the real numbers  $\varepsilon_1$  and  $\varepsilon_2$  vanish as  $\Delta x, \Delta y \rightarrow 0$ . Using a similar formula for  $v(x, y)$ , we have

$$\begin{aligned}\Delta f &= f(z_0 + \Delta z) - f(z_0) = \Delta u + i\Delta v \\ &= \frac{\partial u}{\partial x} \Delta x + \frac{\partial u}{\partial y} \Delta y + \varepsilon_1 \Delta x + \varepsilon_2 \Delta y \\ &\quad + i \left( \frac{\partial v}{\partial x} \Delta x + \frac{\partial v}{\partial y} \Delta y + \varepsilon_3 \Delta x + \varepsilon_4 \Delta y \right).\end{aligned}$$

Using the CR equations that are supposed to hold at the point  $(x_0, y_0)$  from assumption **(ii)** above gives us

$$\Delta f = \left( \frac{\partial u}{\partial x} + i \frac{\partial v}{\partial x} \right) (\Delta x + i\Delta y) + \Delta x(\varepsilon_1 + i\varepsilon_3) + \Delta y(\varepsilon_2 + i\varepsilon_4).$$

Dividing the both sides by  $\Delta z = \Delta x + i\Delta y$  yields

$$\frac{\Delta f}{\Delta z} = \frac{\partial u}{\partial x} + i \frac{\partial v}{\partial x} + (\varepsilon_1 + i\varepsilon_3) \frac{\Delta x}{\Delta z} + (\varepsilon_2 + i\varepsilon_4) \frac{\Delta y}{\Delta z}.\tag{7.6}$$

Since  $|\Delta z| = \sqrt{(\Delta x)^2 + (\Delta y)^2}$ , we have

$$|\Delta x| \leq |\Delta z| \quad \text{and} \quad |\Delta y| \leq |\Delta z|,$$

so that

$$\left| \frac{\Delta x}{\Delta z} \right| \leq 1 \quad \text{and} \quad \left| \frac{\Delta y}{\Delta z} \right| \leq 1.\tag{7.7}$$

Hence, it follows from (7.7) that the last two terms in (7.6) tend to zero with  $\Delta z \rightarrow 0$  because  $\lim_{\Delta z \rightarrow 0} \varepsilon_n = 0$  ( $1 \leq n \leq 4$ ). As a result, the limit

$$\lim_{\Delta z \rightarrow 0} \frac{\Delta f}{\Delta z} = \frac{\partial u}{\partial x} + i \frac{\partial v}{\partial x}\tag{7.8}$$

is independent of the path of  $\Delta z \rightarrow 0$ , so the derivative  $f'(z_0)$  exists. We thus have verified that  $f(z)$  is differentiable at  $z_0$  if conditions **(i)** and **(ii)** are satisfied. This completes the proof of the analyticity of  $f(z)$ . ♣

*Examples* 1. Regarding the function

$$f(z) = z^2 = (x^2 - y^2) + i(2xy) \equiv u + iv, \quad (7.9)$$

we have

$$\frac{\partial u}{\partial x} = 2x = \frac{\partial v}{\partial y}, \quad \text{and} \quad \frac{\partial v}{\partial x} = 2y = -\frac{\partial u}{\partial y}. \quad (7.10)$$

These equations mean that everywhere in the complex plane the CR relations hold and the partial derivatives are continuous. Hence, the function (7.9) is analytic in the entire complex plane. Such analytic functions are called **entire functions**.

2. We saw in Sect. 7.1.1 that the function  $f(z) = |z|^2 = x^2 + y^2$  is not analytic anywhere since it is differentiable only at the origin. In fact, it yields

$$\frac{\partial u}{\partial x} = 2x, \quad \frac{\partial u}{\partial y} = 2y, \quad \frac{\partial v}{\partial x} = \frac{\partial u}{\partial y} = 0,$$

which satisfy the CR relations only at the origin.

#### 7.1.4 Harmonic Functions

The CR relations immediately provide one remarkable result that points to connections with physics. Provided that the CR relations hold in a region, we set

$$\frac{\partial}{\partial x} \frac{\partial u}{\partial x} = \frac{\partial}{\partial x} \frac{\partial v}{\partial y} = \frac{\partial}{\partial y} \frac{\partial v}{\partial x} = -\frac{\partial}{\partial y} \frac{\partial u}{\partial y}. \quad (7.11)$$

Here we assume the continuity of the second-order partial derivatives of  $u(x, y)$  and  $v(x, y)$ , which allows us to interchange the orders of differentiation in the mixed partial derivatives in (7.11). (This qualification, however, can be dropped since the second-order partial derivatives of an analytic function are necessarily continuous as we prove later.) Equation (7.11) yields the **Laplace equation**:

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \nabla^2 u = 0.$$

In the same way, it follows that

$$\nabla^2 v = 0.$$

Thus we set the following theorem:

♠ **Theorem:**

Each of the real and imaginary parts of analytic functions satisfies the two-dimensional Laplace equation.

Any function  $\phi$  satisfying  $\nabla^2 \phi = 0$  is called an **harmonic function**. Accordingly, if  $f = u + iv$  is an analytic function, then  $u$  and  $v$  are called **conjugate**

**harmonic functions** since  $\nabla^2 u = \nabla^2 v = 0$  holds. The fact that real and imaginary components of analytic functions satisfy the Laplace equation plays a crucial role in solving applied second-order partial differential equations. Detail discussions on this point are presented in Sect. 9.4.3.

### 7.1.5 Geometric Interpretation of Analyticity

To gain in-depth insight into the nature of analytic functions, we reveal the geometric meaning of “**analyticity**.” We know that the analyticity of  $f(z)$  within a domain  $D$  ensures the existence of the derivative  $f'(z) = df/dz$  defined by

$$f'(z) = \lim_{h \rightarrow 0} \frac{f(z+h) - f(z)}{h}.$$

This suggests that at a point  $z_0$  within  $D$ ,

$$f(z_0 + h) - f(z_0) \simeq f'(z_0)h \quad (7.12)$$

for an arbitrary complex number  $h$  the magnitude  $|h|$  is sufficiently small.

Let us consider the geometrical meaning of (7.12). For the discussion to be concrete, we assume, for the moment, that the derivative  $f'(z)$  takes the values

$$f'(z_0) = 1 + i \quad \text{and} \quad f'(z_1) = \frac{-1 + \sqrt{3}i}{2}$$

at the points  $z_0$  and  $z_1$  in  $D$ . It then follows that

$$f'(z_0)h = (1 + i)h = \sqrt{2} \left( \frac{1}{\sqrt{2}} + \frac{i}{\sqrt{2}} \right) h = \sqrt{2}e^{i\pi/4}h, \quad (7.13)$$

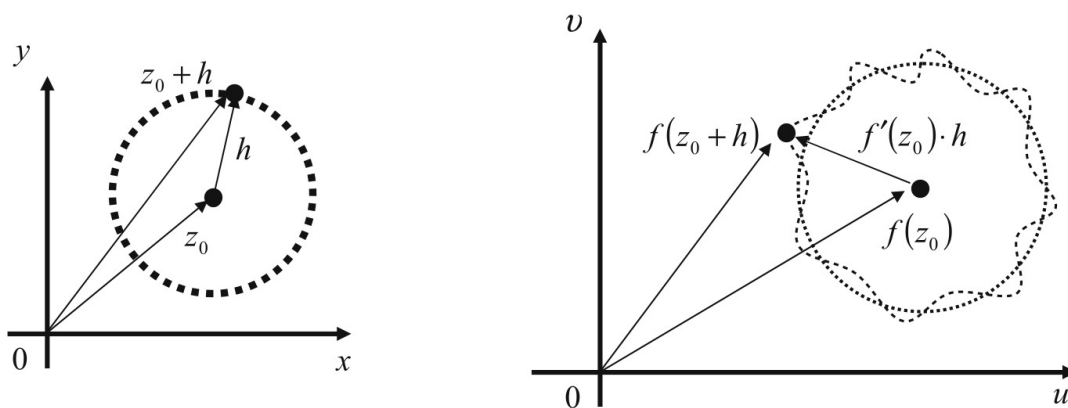
where  $h = |h|(\cos \theta + i \sin \theta)$  is a complex number having a certain argument  $\theta$ . Equation (7.13) means that  $f'(z_0)h$  is obtained through the rotation of the vector  $h$  by  $\pi/4$  followed by multiplication by  $\sqrt{2}$ . (Note that any complex number can be regarded as a vector on the two-dimensional complex plane.) Similarly, we have

$$f'(z_1)h = e^{2\pi i/3}h, \quad (7.14)$$

which states that  $f'(z_1)h$  is obtained through the rotation of  $h$  by  $2\pi/3$ . The processes are schematically illustrated in Fig. 7.3. The vector  $h$  is depicted by thin arrows and the corresponding vectors  $f'(z)h$  by thick arrows. Noteworthy is that the magnitude  $|f'(z)h|$  at both  $z_0$  and  $z_1$ , is invariant no matter what direction the vector  $h$  takes; indeed it follows from (7.13) and (7.14) that

$$|f'(z_0)h| = \sqrt{2}|h| \quad \text{and} \quad |f'(z_1)h| = |h|.$$

Hence, when the direction of  $h$  is shifted by increasing  $\theta$ ,  $|f'(z)h|$  remains unchanged so that the front edge of the vector  $f'(z)h$  moves along a circle centered at the origin.



**Fig. 7.3.** Illustration of analyticity of  $f(z)$  at  $z_0$ . An infinitesimal circle on the  $z$ -plane centered at an analytic point is mapped to a circle on the  $w$ -plane with slight modulation

Now we go back to (7.12), which says that if  $f(z)$  is analytic at  $z_0$ , the acquired vectors  $f'(z_0)h$  given above are almost equal to the vectors  $f(z_0 + h) - f(z_0)$ . This implies that the magnitude  $|f(z_0 + h) - f(z_0)|$  is almost invariant to the change in the direction of  $h$  characterized by  $\theta$ . Thus as  $\theta$  increases, the front edge of  $f(z_0 + h) - f(z_0)$  should trace a circle centered at the origin. (To be precise, the radius may be subjected to a slight fluctuation, as shown in Fig. 7.3, owing to contributions from higher-order terms than  $h^2$ .) In other words, since  $f(z_0)$  is fixed, an increase in  $\theta$  from 0 to  $2\pi$  results in movement of  $f(z_0 + h)$  along the circle centered at  $f(z_0)$ . This means that for analytic functions  $f(z)$ , the change in the magnitude of  $f$  for an infinitesimal change in  $z$  is isotropic. This isotropy is the geometric interpretation of the analyticity of  $f(z)$ .

Better understanding can be attained by considering the case of nonanalytic functions. Let us use the same argument for the function

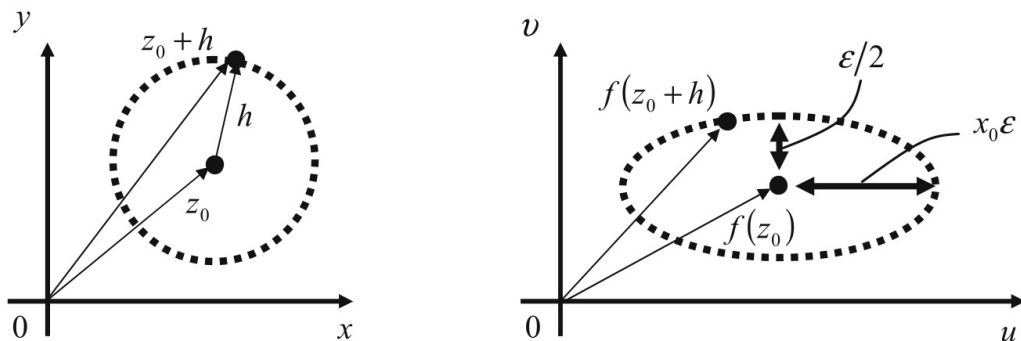
$$f(z) = x^2 + iy, \quad (7.15)$$

where  $u = x^2$  and  $v = y$ . This function is not analytic, since it does not satisfy the CR relations. Indeed,

$$\frac{\partial u}{\partial x} = 2x \neq 1 = \frac{\partial v}{\partial y}$$

except at  $x = 1/2$ . For such a nonanalytic function, the isotropy regarding the magnitude of the difference  $|f(z + h) - f(z)|$  for infinitesimal  $h$  breaks down, as is shown below. Once we set  $h = |h|(\cos \theta + i \sin \theta)$  with  $|h| = \varepsilon = \text{const}$ , we have

$$\begin{aligned} f(z_0 + h) &= (x_0 + \varepsilon \cos \theta)^2 + i(y_0 + \varepsilon \sin \theta) \\ &\simeq x_0^2 + 2\varepsilon \cos \theta \cdot x_0 + iy_0 + i\varepsilon \sin \theta \\ &= f(z_0) + 2\varepsilon \cos \theta \cdot x_0 + i\varepsilon \sin \theta, \end{aligned} \quad (7.16)$$



**Fig. 7.4.** Schematic illustration of nonanalyticity. When  $f(z)$  is not analytic at  $z = z_0$ , then an infinitesimal circle centered at  $z_0$  is mapped to an ellipse so the isotropy breaks out

up to the order of  $\varepsilon$ . Equation (7.16) indicates that when  $\theta$  increases, the front edge of the vector  $f(z_0 + h)$  moves along an ellipse that has a major axis of  $2x_0\varepsilon$  and a minor axis  $\varepsilon$  (see Fig. 7.4). That is, the magnitude  $|f(z_0 + h) - f(z_0)|$  is no longer isotropic, but depends on the direction of  $h$  (except for the particular case of  $x_0 = 1/2$ ).

## Exercises

1. Show that  $f(z)$  is continuous at  $z_0$  if it is analytic there.

**Solution:** From the identity, we have

$$f(z) - f(z_0) = f(z_0 + \Delta z) - f(z_0) = \Delta z \cdot \frac{f(z_0 + \Delta z) - f(z_0)}{\Delta z}$$

and with the definition  $\Delta z = z - z_0$ , we set

$$\lim_{\Delta z \rightarrow 0} [f(z_0 + \Delta z) - f(z_0)] = \left( \lim_{\Delta z \rightarrow 0} \Delta z \right) f'(z_0) = 0.$$

Moreover, if we write  $f(z) = u(z) + iv(z)$ , it follows that  $u(z)$  and  $v(z)$  are both continuous. ♣

2. Express the Cauchy–Riemann relations in polar coordinates  $(r, \theta)$ .

**Solution:** By imposing  $z = x + iy = re^{i\theta}$ , we transform the partial derivatives in terms of  $x$  into  $\partial/\partial x = (\partial r/\partial x)(\partial/\partial r) + (\partial\theta/\partial x)(\partial/\partial\theta)$ . After some algebra, we obtain  $\partial/\partial x = \cos\theta(\partial/\partial r) - (\sin\theta/r)(\partial/\partial\theta)$ , which, together with the same procedure with respect to  $\partial/\partial y$ , yields the polar form of the CR relations as

$$\frac{\partial u}{\partial r} = \frac{1}{r} \frac{\partial v}{\partial \theta}, \quad \frac{\partial u}{\partial \theta} = -r \frac{\partial v}{\partial r}.$$

Their abbreviated forms read  $u_r = v_\theta/r$  and  $u_\theta = -rv_r$ . ♣

3. If  $f(z)$  is analytic in a region  $D$  and if  $|f(z)|$  is constant there, then  $f(z)$  is constant. Prove it.

**Solution:** If  $|f| = 0$ , the proof is immediate. Otherwise we have

$$u^2 + v^2 \equiv c \neq 0. \quad (7.17)$$

Taking the partial derivatives with respect to  $x$  and  $y$ , we have  $uu_x + vv_x \equiv 0$  and  $uu_y + vv_y \equiv 0$ . Using the CR relations, we obtain  $uu_x - vv_y \equiv 0$  and  $vu_x + uu_y \equiv 0$ , so that

$$(u^2 + v^2)u_x \equiv 0. \quad (7.18)$$

From (7.17) and (7.18), and from the CR relations, we conclude that  $u_x = v_y \equiv 0$ . We can obtain  $u_y = v_x \equiv 0$  in a similar manner.

Therefore,  $f$  is constant. ♣

4. Let  $\phi(x, y)$  and  $\psi(x, y)$  be harmonic functions in a domain  $D$ . Show that if we set  $u = \phi_y - \psi_x$  and  $v = \phi_x - \psi_y$ , the function  $f(z) = u + iv$  with the variable  $z = x + iy$  becomes analytic in  $D$ .

**Solution:** It follows that  $u_x - v_y = (\phi_{yx} - \psi_{xx}) - (\phi_{xy} - \psi_{yy}) = -\nabla^2\psi$ , where  $\psi_{yx} = \psi_{xy}$  was used. Since  $\nabla^2\psi = 0$ , we have  $u_x = v_y$ . Similarly, we obtain  $u_y = -v_x$ . Hence,  $u$  and  $v$  satisfy the CR relations in  $D$ , which indicates the analyticity of  $f$  on  $D$ . ♣

## 7.2 Complex Integrations

### 7.2.1 Integration of Complex Functions

We now turn to the integration of functions  $f(z)$  with respect to a complex variable  $z$ . The theory of integration in the complex plane is just the theory of the line integral as defined by

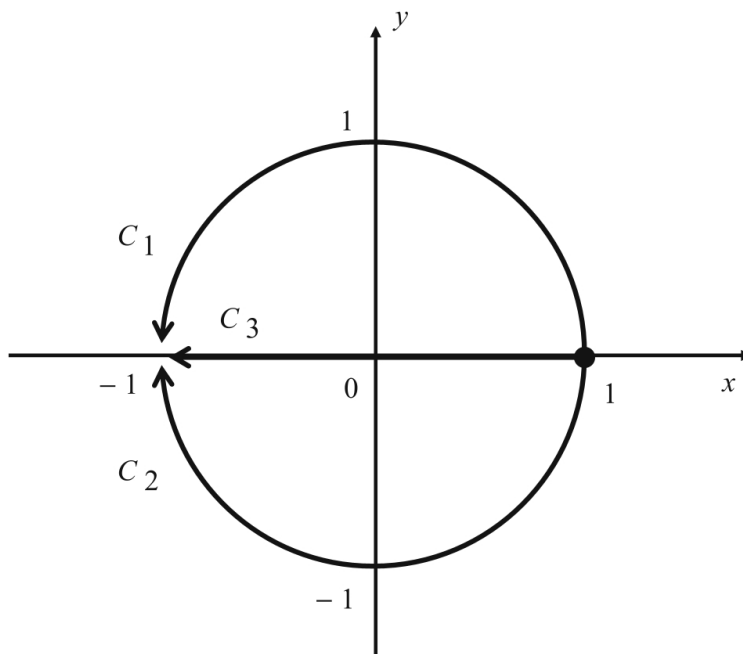
$$\int_{\alpha_1}^{\alpha_2} f(z) dz = \lim_{N \rightarrow \infty, \Delta z_i \rightarrow 0} \sum_{i=1}^N f(z_i) \Delta z_i.$$

Here  $(\Delta z_i)$  is a sequence of small segments situated at  $z_i$  of the curve that connects the complex number  $\alpha_1$  to the other number  $\alpha_2$  in the  $z$ -plane. Since there are infinitely many choices for connecting  $\alpha_1$  to  $\alpha_2$ , it is possible to obtain different values for the integral for different paths.

*Examples* Assume the contour integral

$$I = \oint_{C_i} z^* dz$$

from  $z = 1$  to  $z = -1$  along the three paths (see Fig. 7.5): **(i)** the unit circle centered at the origin in the counterclockwise direction, designated by  $C_1$ ; **(ii)** that in the clockwise direction, denoted by  $C_2$ ; and **(iii)** the real axis,  $C_3$ .



**Fig. 7.5.** Three paths.  $C_1$ ,  $C_2$ , and  $C_3$

- (i) The values of  $z$  on the circle are given by  $z = e^{i\theta}$ , so  $dz = ie^{i\theta} d\theta$ . Thus,

$$I(C_1) = \oint_{C_1} z^* dz = \int_0^\pi e^{-i\theta} ie^{i\theta} d\theta = \pi i. \quad (7.19)$$

- (ii) In a similar manner as in (i), we have

$$I(C_2) = \oint_{C_2} z^* dz = \int_0^{-\pi} e^{-i\theta} ie^{i\theta} d\theta = -\pi i. \quad (7.20)$$

- (iii) On the real axis,  $z = x$  and  $dz = dx$  so that

$$I(C_3) = \oint_{C_3} z^* dz = \int_1^{-1} x dx = -2. \quad (7.21)$$

In general, complex integrals on the path  $C$  possess the following property:

♠ **Darboux's inequality:**

Contour integrals on a path  $C$  satisfy the relation

$$\left| \int_C f(z) dz \right| \leq ML, \quad (7.22)$$

where  $M = \max |f(z)|$  on  $C$  and  $L$  is the length of  $C$ .

This property is very useful because in working with complex line integrals it is often necessary to establish upper bounds on their absolute values.

*Proof* Recall the original definition of complex integrals:

$$\int_C f(z)dz = \lim_{n \rightarrow \infty} \sum_{k=1}^n f(z_k) \Delta z_k.$$

It follows that

$$\left| \sum_{k=1}^n f(z_k) \Delta z_k \right| \leq \sum_{k=1}^n |f(z_k)| |\Delta z_k| \leq M \sum_{k=1}^n |\Delta z_k| \leq ML,$$

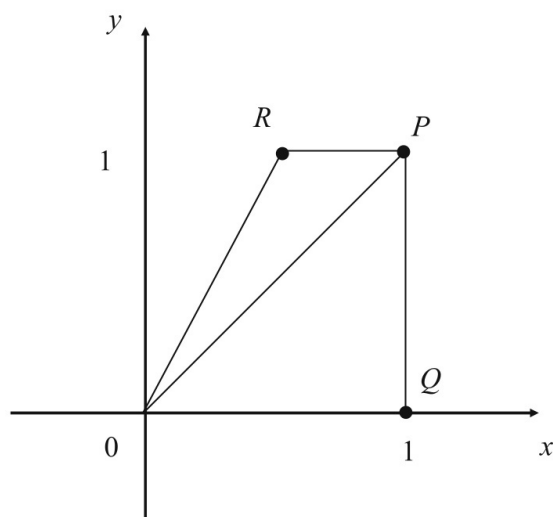
where we have used the facts that  $|f(z)| \leq M$  for all points  $z$  on  $C$ , that the  $\sum |\Delta z_k|$  represents the sum of all the chord lengths joining  $z_{k-1}$  and  $z_k$ , and that this sum is not greater than the length of  $C$ . Taking the limit of both sides, we obtain the desired inequality (7.22). ♣

### 7.2.2 Cauchy Theorem

We are now in a position to proceed with the key theorem in the theory of functions of a complex variable. Consider the complex integral

$$I(C_i) = \oint_{C_i} \sin z dz$$

along the closed paths  $C_i$  ( $i = 1, 2, 3$ ) shown in Fig. 7.6: **(a)**  $C_1 = OP$ , **(b)**  $C_2 = OQ + QP$ , **(c)**  $C_3 = OR + RP$ . After some algebra, we obtain



**Fig. 7.6.** Three paths:  $OQP$ ,  $OP$ , and  $ORP$

$$I(C_1) = I(C_2) = I(C_3) = \dots,$$

which gives us the possibility that the integral from  $O$  to  $B$  remains invariant in quantity for our choices of integration paths. Actually, this is entirely true; it depends only on the two endpoints  $O$  and  $B$ . This peculiarity of integration comes from the fact that the integrand  $\sin z$  is analytic on the integration paths in question. (In fact, it is analytic everywhere on the complex plane.) This result can be generalized to the following statement, called **Cauchy's theorem**, which is pivotal in the theory of complex function analysis.

♠ **Cauchy's theorem:**

If  $f(z)$  is analytic within and on a closed contour  $C$ , then

$$\oint_C f(z) dz = 0. \quad (7.23)$$

The somewhat lengthy discussions that are needed for a proof of Cauchy's theorem, are beyond the scope of this textbook, but two immediate corollaries of the theorem are listed below.

♠ **Path independence:**

If  $f(z)$  is analytic in the region  $R$  and if contours  $C_1$  and  $C_2$  lie in  $R$  and have the same endpoints, then

$$\int_{C_1} f dz = \int_{C_2} f dz.$$

The proof readily follows by applying Cauchy's theorem to the closed contour consisting of  $C_2$  and  $-C_1$  as shown in Fig. 7.7:

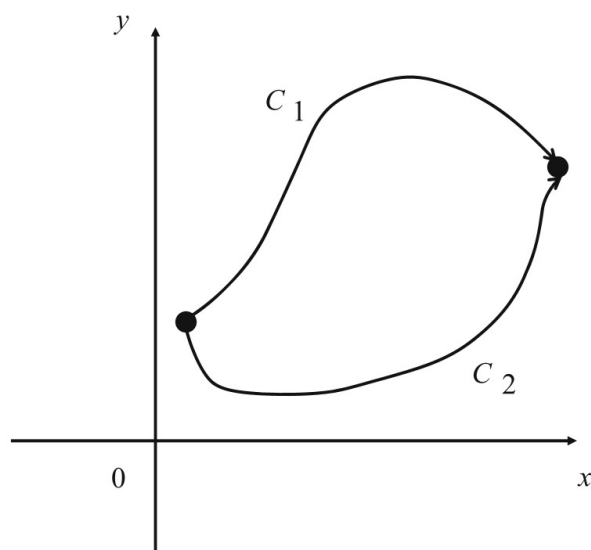
$$\int_{C_2} + \int_{-C_1} = 0 \Rightarrow \int_{C_2} = - \int_{-C_1} = \int_{C_1}.$$

Intuitively, the symbol  $-C$  denotes the contour  $C$  traced in the opposite direction. A discussion on the path independence follows the theorem below.

♠ **Uniqueness of the integral:**

If  $f(z)$  is analytic within a region bounded by a closed contour  $C$ , then the integration  $\int_{z_1}^{z_2} f(z) dz$  along any contour within  $C$  depends only on  $z_1$  and  $z_2$ .

This theorem states that an analytic function  $f(z)$  has a unique integral not only a unique derivative. From a practical viewpoint, this theorem is frequently

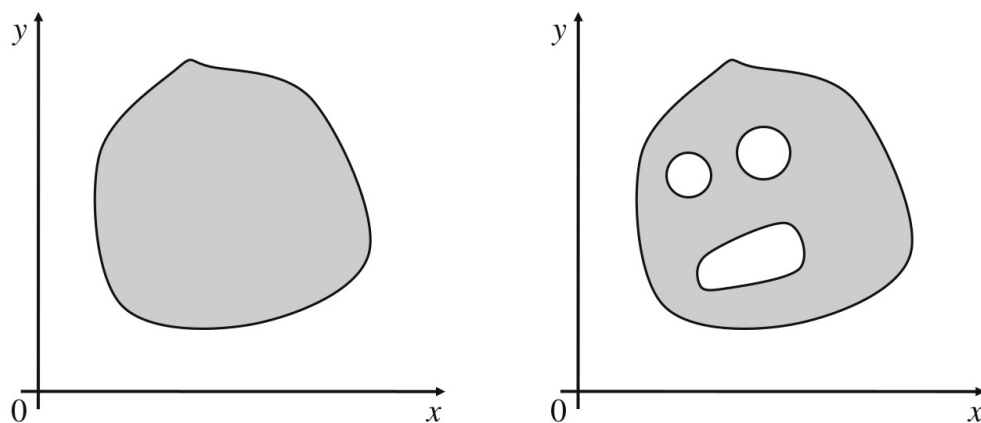


**Fig. 7.7.** Two integration paths:  $C_1$  and  $C_2 = -C_1$

used in the evaluation of contour integrals, since it allows us to choose an appropriate contour.

**Remark.** When integrating along a closed contour, we agree to move along the contour in such a way that the enclosed region lies to our left. An integration that follows this convention is called integration in the *positive* sense. Integration performed in the opposite direction acquires a minus sign.

### 7.2.3 Integrations on a Multiply Connected Region



**Fig. 7.8.** *Left:* A simply connected region. *Right:* A multiply connected region

We should note that Cauchy's theorem applies in a direct way only to simply connected regions. A region  $R$  is said to be **simply connected** if every

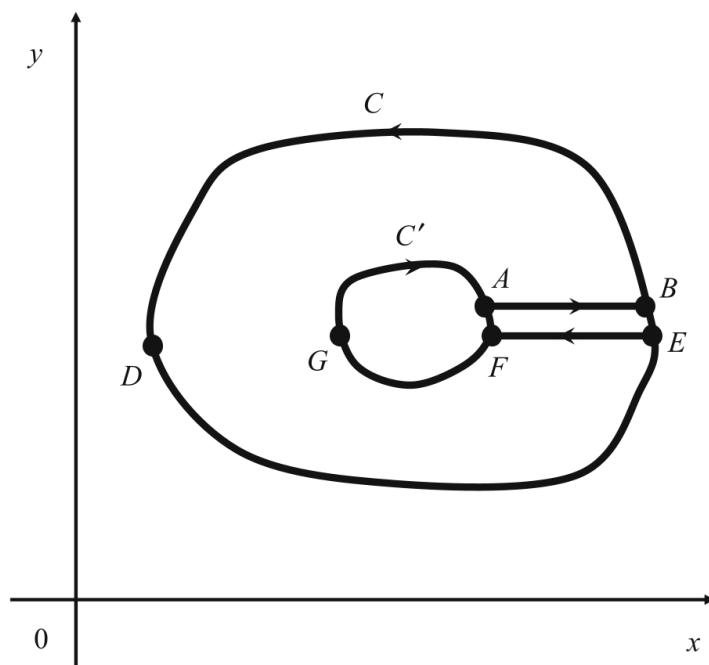
closed curve in  $R$  can be continuously contracted into a point without leaving  $R$ . Otherwise, it is said to be **multiply connected**; (see Fig. 7.8). The physical reason for this restriction is easy to find. The important fact is that Cauchy's theorem is a restatement that no singular point is included within the region bounded by the contour  $C$ . If the region  $R$  bounded by  $C$  is multiply connected, it becomes possible to put on singular points *within the closed contour*  $C$  but surely *outside the region*  $R$  in question. In this case, Cauchy's theorem no longer holds even though the integrand  $f(z)$  is analytic everywhere in the region.

Nevertheless, there is still a way to apply Cauchy's theorem to multiply connected regions, which is based on allowing the deformation of contours as described below.

Suppose that  $f(z)$  is analytic in the region that lies between two closed contours  $C$  and  $C'$ , where  $C$  encloses  $C'$ . Draw two lines  $AB$  and  $EF$  close together, so as to connect the two contours. Then  $ABDEFGA$  described as shown in Fig. 7.9 is a closed contour, which we denote by  $S$  and  $f(z)$  is analytic within it. Then, we have

$$\oint_S f(z)dz = 0.$$

Now let the lines  $AB$  and  $FE$  approach infinitely close to one another. The contribution from the part  $BDE$  tends toward the integral around  $C$  in the positive (i.e., counterclockwise) direction. Similarly, the contribution from  $FGA$  tends toward that around  $C'$  in the negative (clockwise) direction, thus minus that around  $C'$  in the positive direction. The contributions from  $AB$



**Fig. 7.9.** Closed contour of  $ABDEFGA$  that consists of  $C$  and  $C'$

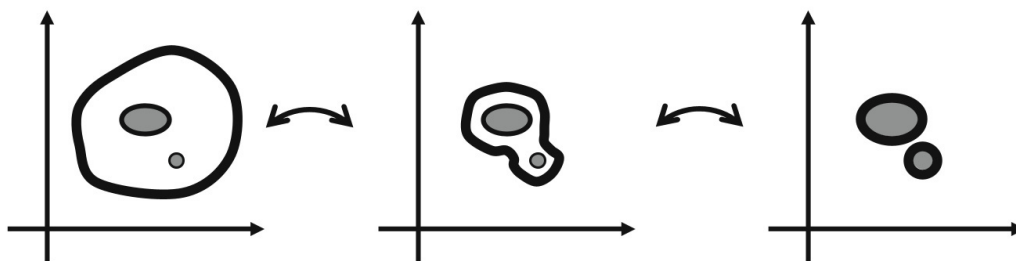
and  $EF$  approach equal and opposite values since they ultimately become the same path described in opposite directions. We thus come to the conclusion that

$$\oint_C f(z)dz = \oint_{C'} f(z)dz.$$

This means that, *if a function is analytic between two contours, its integrals around both contours have the same value.*

**Remark.** There is an immediate extension to the case where  $C$  encloses several closed paths  $C_1, C_2, \dots$ , all external to one another. Because of Cauchy's theorem, an integration contour can be moved across any region of the complex plane over which the integrand is analytic without changing the value of the integral. It cannot be moved across a hole (the shaded area) or a singularity (the dot), but it can be made to collapse around either, as shown in Fig. 7.10. As a result, an integration contour  $C$  enclosing  $n$  holes or singularities can be replaced by  $n$  separated closed contours  $C_i$ , each enclosing a hole or a singularity as given by

$$\oint_C f(z)dz = \sum_{i=1}^n \oint_{C_i} f(z)dz.$$



**Fig. 7.10.** Collapse of an integration path onto the boundaries of a hole (a large shaded region) and singularity (a small shaded dot)

### 7.2.4 Primitive Functions

Here is a definition of the primitive function of a complex function.

#### ♠ Primitive function:

Let  $f(z)$  be a function that is continuous in a domain  $D$  and has the property  $\oint_C f(z)dz = 0$  for every closed path  $C$  in  $D$ . Then, the primitive function  $F(z)$  of  $f(z)$  is defined by

$$F(z) = \int_{z_0}^z f(z')dz' \quad (z_0, z \in D),$$

which is analytic in  $D$  with the derivative

$$\frac{dF(z)}{dz} = f(z).$$

*Proof* Consider the differential

$$F(z + \Delta z) - F(z) = \int_{z_0}^{z+\Delta z} f(z') dz' - \int_{z_0}^z f(z') dz' = \int_z^{z+\Delta z} f(z') dz', \quad (7.24)$$

where we make use of the path-independence property. If we write

$$\begin{aligned} \int_z^{z+\Delta z} f(z') dz' &= f(z) \int_z^{z+\Delta z} dz' + \int_z^{z+\Delta z} [f(z') - f(z)] dz' \\ &= f(z) \Delta z + \int_z^{z+\Delta z} [f(z') - f(z)] dz', \end{aligned}$$

then (7.24) becomes

$$F(z + \Delta z) - F(z) - f(z) \Delta z = \int_z^{z+\Delta z} [f(z') - f(z)] dz'. \quad (7.25)$$

Since  $f(z)$  is continuous, corresponding to an arbitrary small positive number  $\varepsilon$ , there is a number  $\delta$  such that

$$|z - z'| < \delta \Rightarrow |f(z) - f(z')| < \varepsilon.$$

Now choose  $|\Delta z| < \delta$ , which ensures  $|z - z'| < \delta$  for  $z'$  on the path  $C$  in question. Therefore, we have

$$\left| \int_z^{z+\Delta z} [f(z') - f(z)] dz' \right| \leq \int_z^{z+\Delta z} |f(z') - f(z)| |dz'| < \varepsilon |\Delta z|$$

and (7.25) can be written as

$$\left| \frac{F(z + \Delta z) - F(z)}{\Delta z} - f(z) \right| < \varepsilon \text{ for } |\Delta z| < \delta.$$

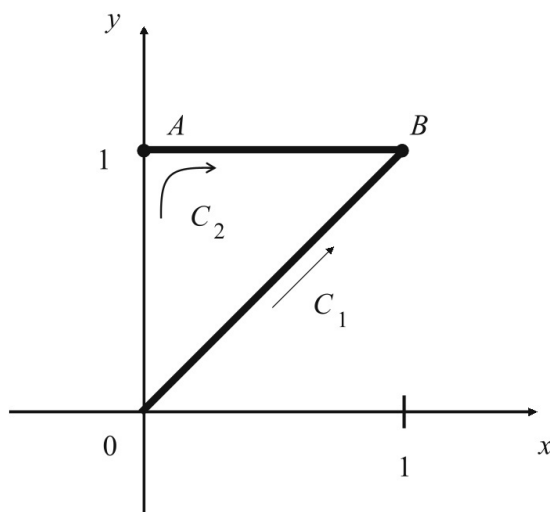
Since  $\varepsilon$  can be arbitrarily small, we conclude that

$$\lim_{\Delta z \rightarrow 0} \frac{F(z + \Delta z) - F(z)}{\Delta z} = f(z),$$

or equivalently,

$$\frac{dF(z)}{dz} = f(z).$$

This result is obtained for any point in  $D$ , so  $F(z)$  is analytic in  $D$ . ♣



**Fig. 7.11.** Integration paths used in Exercise 1

## Exercises

1. Evaluate the integral

$$I(C) = \int_C \sin z dz \quad (7.26)$$

along the two paths shown in Fig. 7.11: **(a)**  $C_1 = OB$ , **(b)**  $C_2 = OA + AB$ .

**Solution:** Since  $\sin z = \sin(x + iy) = \cosh y \sin x + i \sinh y \cos x$  and  $dz = dx + i dy$ , we can divide (7.26) into real and imaginary parts as

$$\begin{aligned} I(C) &= \int_C (\cosh y \sin x dx - \sinh y \cos x dy) \\ &\quad + \int_C (\cosh y \sin x dy + \sinh y \cos x dx). \end{aligned}$$

Noting  $x = y$  along the curve  $C_1$ , we have

$$\begin{aligned} I(C_1) &= (1+i) \int_0^1 \cosh x \sin x dx - (1-i) \int_0^1 \sinh x \cos x dx \\ &= [\cosh x \cos x]_0^1 + [\sinh x \sin x]_0^1 \\ &= (1 - \cosh 1 \cos 1) + i(\sinh 1 \sin 1), \end{aligned} \quad (7.27)$$

where we employ partial integrations. Next we evaluate  $I$  along  $C_2$ . Along the path from  $O$  to  $A$ ,  $x = 0$  and  $dx = 0$ , and along the path from  $A$  to  $B$ ,  $y = 1$  and  $dy = 0$ . Therefore,

$$\begin{aligned} I(C_2) &= \int_{C_2} \sin z dz \\ &= - \int_0^1 \sinh y dy + \int_0^1 \cosh x \sin x dx + i \int_0^1 \sinh x \cos x dx \\ &= (1 - \cosh 1 \cos 1) + i(\sinh 1 \sin 1). \end{aligned}$$

Observe that  $I(C_1) = I(C_2)$ . ♣

2. Set  $C : |z| = r$ , and calculate the following integrals:

$$(i) \left| \oint_C \frac{dz}{z} \right|, \quad (ii) \oint_C \frac{dz}{|z|}, \quad (iii) \oint_C \frac{|dz|}{z}.$$

**Solution:** Let  $z = re^{i\theta}$ , which yields  $dz = ire^{i\theta} d\theta$  and  $|dz| = rd\theta$ . Hence, we have the results: (i)  $|\oint_C dz/z| = |\int_0^{2\pi} (ire^{i\theta})/(re^{i\theta}) d\theta| = 2\pi$ , (ii)  $\oint_C dz/|z| = \int_0^{2\pi} (ire^{i\theta})/rd\theta = 0$ , (iii)  $\oint_C |dz|/z = \int_0^{2\pi} r/(re^{i\theta}) d\theta = 0$ . ♣

3. Let  $f(z)$  be analytic on a unit circle  $D$  about the origin. For any two points  $z_1$  and  $z_2$  on  $D$ , there exists two points  $\xi_1$  and  $\xi_2$  on the line segment  $[z_1, z_2]$  that satisfy the relation

$$f(z_2) - f(z_1) = \{\operatorname{Re}[f'(\xi_1)] + i\operatorname{Im}[f'(\xi_2)]\}(z_2 - z_1). \quad (7.28)$$

Prove it. (This is a generalization of the **mean value theorem** that is valid for real functions.)

**Solution:** From assumptions, we have

$$\begin{aligned} f(z_2) - f(z_1) &= \int_{z_1}^{z_2} f'(z) dz = (z_2 - z_1) \int_0^1 f'[z_1 + t(z_2 - z_1)] dt \\ &= (z_2 - z_1) \left\{ \int_0^1 \operatorname{Re}[f'(z_1 + t(z_2 - z_1))] dt \right. \\ &\quad \left. + i \int_0^1 \operatorname{Im}[f'(z_1 + t(z_2 - z_1))] dt \right\}. \end{aligned} \quad (7.29)$$

Note that the integrals in the last line are both real. Hence, they satisfy the mean value theorem for integrals of real-valued functions  $g(t)$  that are expressed by

$$\int_0^1 g[z_1 + t(z_2 - z_1)] dt = g[z_1 + c(z_2 - z_1)] \quad \text{when } 0 < c < 1.$$

Setting  $\xi_k = z_1 + c_k(z_2 - z_1)$  with  $0 < c_k < 1$  ( $k = 1, 2$ ), we get the desired equation (7.28). ♣

## 7.3 Cauchy Integral Formula and Related Theorem

### 7.3.1 Cauchy Integral Formula

We now turn to the famous integral formula that is the chief tool in the application of the theory of analytic functions in physics.

♠ **Cauchy integral formula:**

If  $f(z)$  is analytic within and on a closed contour  $C$ , we have

$$\oint_C \frac{f(z)}{z-a} dz = \begin{cases} 2\pi i f(a) & \text{if } a \text{ is interior to } C, \\ 0 & \text{if } a \text{ is exterior to } C. \end{cases} \quad (7.30)$$

*Proof* The latter case is trivial; when  $z = a$  is exterior to  $C$ , the integrand in (7.30) becomes analytic within  $C$  so that we have at once  $\oint [f(z)/(z-a)] dz = 0$  by virtue of the Cauchy theorem. Hence, we consider below only the case where  $z = a$  is within  $C$ .

Suppose that the integral

$$J(a) = \oint_C \frac{f(z)}{z-a} dz \quad (7.31)$$

around a closed contour  $C$  within and on which  $f(z)$  is analytic. In view of the discussion in Sect. 7.2.3, the contour  $C$  may be deformed into a small circle of radius  $r$  about the point  $a$ . Accordingly, the variable  $z$  is expressed by  $z = a + re^{i\theta}$ .

Now, we rewrite (7.31) as

$$J(a) = f(a) \oint_C \frac{dz}{z-a} + \oint_C \frac{f(z) - f(a)}{z-a} dz. \quad (7.32)$$

The first integral on the right-hand side becomes

$$\oint_C \frac{dz}{z-a} = \int_0^{2\pi} \frac{ire^{i\theta}}{re^{i\theta}} d\theta = 2\pi i. \quad (7.33)$$

Hence, (7.30) is confirmed if the second integral of (7.32) vanishes for some choice of the radius  $r$  of the circle  $C$ . To show this, we note the continuity of  $f(z)$  at  $a$ , which tells us that for all  $\varepsilon > 0$  there exists an appropriate quantity  $\delta$  such that

$$|z-a| < \delta \Rightarrow |f(z) - f(a)| < \varepsilon.$$

This implies that for any arbitrarily small  $\varepsilon$ , we can find  $r = |z-a|$  that satisfies the relation

$$\left| \oint_C \frac{f(z) - f(a)}{z-a} dz \right| \leq \oint_C \frac{|f(z) - f(a)|}{|z-a|} |dz| < \frac{\varepsilon}{\delta} 2\pi\delta = 2\pi\varepsilon. \quad (7.34)$$

Thus by taking  $r$  small enough, but still greater than zero, the absolute value of the integral can be made smaller than any preassigned number. From (7.32 to 7.34), we obtain the desired equation:

$$\oint \frac{f(z)}{z-a} dz = 2\pi i f(a) \text{ if } a \text{ is within } C. \quad \clubsuit \quad (7.35)$$

**Remark.** If  $a$  is a point located just on the contour  $C$ , the integral (7.30) will have the **principal value integral** (see Sect. 9.4.1).

The Cauchy integral formula gives us another hint by which to comprehend the rigid structure of analytic functions: If a function is analytic within and on a closed contour  $C$ , its value at every point inside  $C$  is determined by its values on the bounding curve  $C$ .

### 7.3.2 Goursat Formula

A remarkable consequence of the Cauchy's integral formula is the fact that, when  $f(z)$  is analytic at  $z = a$ , all of its derivatives are also analytic. Furthermore, the region of analyticity for those derivatives is identical with that of  $f(z)$ . To prove the theorem, we use the integral representation (7.35) to evaluate the derivative,

$$\begin{aligned} & 2\pi i f'(a) \\ &= 2\pi i \lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h} = \lim_{h \rightarrow 0} \left[ \frac{1}{h} \oint f(z) \left( \frac{1}{z-a-h} - \frac{1}{z-a} \right) dz \right] \\ &= \lim_{h \rightarrow 0} \oint \frac{f(z)}{(z-a)(z-a-h)} dz = \oint \frac{f(z)}{(z-a)^2} dz. \end{aligned} \quad (7.36)$$

The last equality in (7.36) is verified from

$$\begin{aligned} & \oint \left[ \frac{f(z)}{(z-a-h)(z-a)} - \frac{f(z)}{(z-a)^2} \right] dz \\ &= h \oint \frac{f(z)}{(z-a)^2(z-a-h)} dz \leq \frac{hML}{b^2(b-|h|)}, \end{aligned} \quad (7.37)$$

where  $M$  is the maximum value of  $|f(z)|$  on the contour,  $L$  is the length of the contour, and  $b$  is the minimum value of  $|z-a|$  on the contour. The right-hand side of the inequality in (7.37) approaches zero as  $h \rightarrow 0$ , so we have

$$\lim_{h \rightarrow 0} \oint \left[ \frac{f(z)}{(z-a)(z-a-h)} - \frac{f(z)}{(z-a)^2} \right] dz = 0,$$

which ensures the equality of the last part of (7.36).

We can continue with the same process to obtain higher derivatives, arriving at the general formula for the  $n$ th derivative of  $f$  at  $z = a$ :

♠ **Goursat formula:**

If  $f(z)$  is analytic within and on a closed contour  $C$ , we have

$$f^{(n)}(a) = \frac{n!}{2\pi i} \oint \frac{f(z)}{(z-a)^{n+1}} dz \quad (n = 0, 1, 2, \dots). \quad (7.38)$$

Note that equation (7.38) guarantees the existence of all the derivatives  $f'(a), f''(a), \dots$  and the analyticity at all  $a$ 's within  $C$ .

**Remark.** The Goursat formula (7.38) is valid only within the contour, and thus gives no information as to the existence of the derivatives *just on* the contour.

### 7.3.3 Absence of Extrema in Analytic Regions

An additional noteworthy fact associated with **Cauchy's integral formula** (7.30) is that it points up the absence of either a maximum or a minimum of an analytic function within a **region of analyticity**.

For example, if  $z = a$  is a point within  $C$ , from (7.30) we see that

$$f(a) = \frac{1}{2\pi} \int_0^{2\pi} f(a + \rho e^{i\phi}) d\phi, \quad (7.39)$$

which means that  $f(a)$  is the arithmetic average of the values of  $f(z)$  on any circle centered at  $a$ . We thus have  $|f(a)| \leq M$ , where  $M$  is the maximum value of  $|f|$  *just on* the circle. (Equality can occur only if  $f$  is constant on the contour.)

The above argument applies to arbitrary points *within* the circle and, further, to a region bounded by any contour  $C$  (not necessary a circle). We thus conclude that the inequality  $|f(z)| \leq M$  holds for all  $z$  within  $C$ , which means that  $|f(z)|$  has no maximum within the region of analyticity.

Similarly, if  $f(z)$  has no zero within  $C$ , then  $1/f(z)$  is an analytic function inside  $C$  and  $|1/f(z)|$  has no maximum within  $C$ , taking its maximum value on  $C$ . Therefore  $|f(z)|$  does not have a minimum within  $C$  but does have one on the contour  $C$ . We thus arrive at the following important theorem.

♠ **Absolute maximum theorem:** If a nonconstant function  $f(z)$  is analytic within and on a closed contour  $C$ , then  $|f(z)|$  can have no maximum within  $C$ .

♠ **Absolute minimum theorem:**

If a nonconstant function  $f(z)$  is analytic within and on a closed contour  $C$ , and if  $f(z) \neq 0$  there, then  $|f(z)|$  can have no minimum within  $C$ .

Accordingly, points at which  $df/dz = 0$  are saddle points, rather than true maxima or minima.

We further observe that the theorems apply not only to  $|f(z)|$  but also to the real and imaginary parts of an analytic function. To see this, we rewrite (7.39) as

$$2\pi f(a) = 2\pi(u_a + iv_a) \quad \text{and} \quad 2\pi f(a) = \int_0^{2\pi} f(x + iy)d\phi = \int_0^{2\pi} (u + iv)d\phi, \quad (7.40)$$

where  $u_a$  and  $v_a$  are the values of  $u(x, y)$  and  $v(x, y)$  at  $z = x + iy = a$ . Equating the last terms of the two equations in (7.40), we obtain

$$u_a = \frac{1}{2\pi} \int_0^{2\pi} u d\phi \quad \text{and} \quad v_a = \frac{1}{2\pi} \int_0^{2\pi} v d\phi,$$

so that  $u_a$  and  $v_a$  are the arithmetic averages of the values of  $u(x, y)$  and  $v(x, y)$ , respectively, on the boundary of the circle. Hence, based on the same reasoning as above, we see that both of  $u$  and  $v$  take on their minimum and maximum values on the boundary curve of a region within which  $f$  is analytic.

### 7.3.4 Liouville Theorem

We saw in the previous discussion that  $|f(z)|$  has its maximum  $M$  on the boundary of the region of analyticity of  $f(z)$ . In certain cases, the maximum of  $|f(z)|$  bounds the absolute value of derivatives  $|f^{(n)}(z)|$ , as stated in the theorem below.

♠ **Cauchy inequality:**

If  $f(z)$  is analytic within and on a circle  $C$  with a radius  $r$ , and  $M(r)$  is the maximum of  $|f(z)|$  on  $C$ , then we have

$$\left| f^{(n)}(z) \right| \leq \frac{n!}{r^n} M(r) \quad \text{within and on } C.$$

This is called the **Cauchy inequality**.

*Proof* Goursat's formula reads

$$f^{(n)}(z_0) = \frac{n!}{2\pi i} \oint_C \frac{f(z)}{(z - z_0)^{n+1}} dz.$$

Take  $|z - z_0| = r$  and use the Darboux inequality to get the desired result:

$$\begin{aligned} \left| f^{(n)}(z_0) \right| &\leq \left| \frac{n!}{2\pi i} \oint_C \frac{|f(z)|}{|z - z_0|^{n+1}} |dz| \right| \leq \frac{n!}{2\pi r^{n+1}} M(r) \cdot 2\pi r \\ &= \frac{n!M(r)}{r^n}. \quad \clubsuit \end{aligned} \quad (7.41)$$

If the  $f(z)$  we have considered is analytic at all points on the complex plane, i.e., if it is an **entire function**, the above result reduces to the following theorem:

♠ **Liouville theorem:**

If  $f(z)$  is an **entire function** and  $|f(z)|$  is bounded for all values of  $z$ , then  $f(z)$  is a constant.

*Proof* Let  $n = 1$  and  $M(r) = M$  in (7.41) to obtain

$$|f'(z_0)| < \frac{M}{r}.$$

Since  $f(z)$  is an entire functions we may take  $r$  as large as we like. Thus we can make  $|f'(z_0)| < \varepsilon$  for any preassigned  $\varepsilon$ . That is,  $|f'(z_0)| = 0$ , which implies that  $f'(z_0) = 0$  for all  $z_0$ , so  $f(z_0) = \text{const.}$  ♣

Liouville's theorem is a very powerful statement about analytic functions over the complex plane. In fact, if we restrict our attention to the real axis, then it becomes possible to find many *real* functions that are entire and bounded but are *not* constant;  $\cos x$  and  $e^{-x^2}$  are cases in point. In contrast, there is no such freedom for complex analytic functions; any analytic function is either not bounded (goes to infinity somewhere on the complex plane) or not entire (is not analytic at some points of the complex plane).

### 7.3.5 Fundamental Theorem of Algebra

The next theorem follows easily from Liouville's theorem and provides a remarkable tie-up between analysis and algebra. In what follows, the points  $z$  at which  $f(z) = 0$  are called the **zeros** of  $f(z)$  or roots of  $f(z)$ .

♠ **Fundamental theorem of algebra:**

Every nonconstant polynomial of degree  $n$  with complex coefficients has  $n$  zeros in the complex plane.

*Proof* Let  $P(z)$  be any polynomial. If  $P(z) \neq 0$  for all  $z$ , then the function  $f(z) = 1/P(z)$  is entire. Moreover, if  $P$  is nonconstant, then  $P \rightarrow \infty$  as  $z \rightarrow \infty$  so that  $f$  is bounded. Hence, in view of Liouville's theorem,  $f$  must be a constant. This result means that  $P$  is also a constant, which is contrary to our assumption that  $P$  is a nonconstant polynomial. We thus conclude that  $P(z)$  has at least one zero in the complex plane.

Furthermore, an induction argument shows that an  $n$ th-degree polynomial has  $n$  zeros (counting multiplicity; see Remark 1 below). If we assume that every  $k$ th-degree polynomial can be written

$$P_k(z) = A(z - \alpha_1) \cdots (z - \alpha_k),$$

it follows that

$$P_{k+1}(z) = A(z - \alpha_0)(z - \alpha_1) \cdots (z - \alpha_k). \quad \clubsuit$$

*Remark.*

1. The point  $\alpha$  is called a **zero of order  $k$**  (or **zero of multiplicity  $k$** ) of the function  $P(z)$  if it reads

$$P(z) = (z - \alpha)^k Q(z),$$

where  $Q(z)$  is a polynomial with  $Q(\alpha) \neq 0$ . Equivalently,  $\alpha$  is a zero of order  $k$  if

$$P(\alpha) = P'(\alpha) = \cdots = P^{(k-1)}(\alpha) = 0 \quad \text{and} \quad P^{(k)}(\alpha) \neq 0.$$

2. It can be shown that if  $f_1(z)$  and  $f_2(z)$  are analytic within and on  $C$  and if  $|f_2(z)| < |f_1(z)| \neq 0$  on  $C$ , then  $f_1(z)$  and  $f_1(z) + f_2(z)$  have the same number of zeros within  $C$ . This is called **Rouché's theorem**, which is verified in Sect. 9.3.4.

### 7.3.6 Morera Theorem

The final important theorem is called Morera's theorem and, in a sense the converse of Cauchy's theorem.

#### ♠ Morera theorem:

Let  $f(z)$  be a continuous function on some domain  $D$  and suppose that

$$\oint_C f(z) dz = 0$$

for every simple closed curve  $C$  in  $D$  whose interior also lies in  $D$ . Then  $f$  is analytic in  $D$ .

*Proof* For some fixed point  $z_0$  in  $D$ , define the function

$$F(z) = \int_{z_0}^z f(z') dz', \quad z \in D,$$

where the path is along the line segment in  $D$  from  $z_0$  to  $z$ . From this, we have

$$\frac{F(z + \Delta z) - F(z)}{\Delta z} = \frac{f(z)}{\Delta z} \int_z^{z+\Delta z} dz' + \frac{1}{\Delta z} \int_z^{z+\Delta z} \{f(z') - f(z)\} dz' = f(z),$$

where Darboux's inequality is used in the second term in the limit  $\Delta z \rightarrow 0$ . As a result, we get

$$F'(z) = f(z),$$

which indicates the existence of the first derivative of  $F(z)$ , so  $F(z)$  is analytic in  $D$  and  $f(z)$  is also analytic. ♣

## Exercises

1. Let  $f(z)$  be analytic within a circle  $D : z = |R|$ , and let it satisfy the relations  $|f(z)| \leq M$  and  $f(0) = 0$ .

(i) Prove that

$$|f(z)| \leq \frac{M}{R} |z| \quad \text{for } z \in D. \quad (7.42)$$

(ii) Prove that the equality in (7.42) holds at  $z = z_0$  if and only if there exists a complex number  $c$  that yields  $|c| = 1$  and

$$f(z_0) = c \frac{M}{R} z. \quad (7.43)$$

Statements (i) and (ii) constitute the **Schwarz lemma**.

**Solution:** (i) Equation in (7.42) holds trivially for  $z = 0$ . For considering the case of  $z \neq 0$ , we specify the circle  $D' : |z| = \rho < R$  and set the function  $g(z) = f(z)/z$ . Since  $g$  is analytic within and on  $D'$ , it follows from the theorem in Sect. 7.3.3 that

$$|g(z)| \leq \max_{z \in D'} |g(z)| \leq \frac{M}{\rho},$$

which means that

$$|f(z)| \leq \frac{M}{\rho} |z| \quad \text{for } z \in D'.$$

By fixing  $z$  within  $D'$  and taking the limit of  $\rho$  to  $R$ , we get to (7.42).

(ii) If the equality in (7.42) holds at some  $z_0 \in D$  except at the origin, we have

$$|g(z_0)| = \frac{M}{R} \geq |g(z)| \text{ for } z \in D.$$

It follows again from the theorem in Sect. 7.3.3 that  $g(z)$  must be constant within  $D$ . Hence, we have

$$g(z) = c \frac{M}{R} \text{ with } |c| = 1.$$

This reduces to the desired result (7.43). ♣

- 2.** Let  $f(z)$  be analytic on a domain  $D$  and  $f(z) \not\equiv 0$ . Show that if  $f(a) = 0$  with  $a \in D$ , then it is always possible to find small  $\rho > 0$  such that

$$0 < |z - a| < \rho \Rightarrow f(z) \neq 0.$$

This means that zeros of  $f(z)$  are necessarily isolated from each other.

**Solution:** Suppose that  $z = a$  is an  $n$ th zero of  $f(z)$ . From the definition of zero of a complex function, there exists an  $n \in \mathbf{N}$  such that

$$p < n \Rightarrow f^{(n)}(a) \neq 0 \text{ and } f^{(p)}(a) = 0.$$

Hence, the Taylor series of  $f(z)$  around  $z = a$  reads

$$f(z) = \sum_{p=0}^{\infty} \frac{f^{(n+p)}(a)}{(n+p)!} (z-a)^{n+p} = (z-a)^n g_n(z),$$

where

$$g_n(z) = \sum_{p=0}^{\infty} \frac{f^{(n+p)}(a)}{(n+p)!} (z-a)^p, \text{ so } g_n(a) = \frac{f^{(n)}(a)}{n!} \neq 0.$$

Since  $g_n(z)$  is analytic at  $a$ , it is continuous there. Thus we can find  $\rho > 0$  such that

$$|z - a| < \rho \Rightarrow |g_n(z) - g_n(a)| < \frac{1}{2} \frac{|f^{(n)}(a)|}{n!}.$$

It follows from the triangular inequality that

$$|g_n(z)| > |g_n(a)| - \frac{1}{2} \frac{|f^{(n)}(a)|}{n!} = \frac{1}{2} \frac{|f^{(n)}(a)|}{n!} > 0.$$

This implies that for our choice of  $\rho$ ,

$$0 < |z - a| < \rho \Rightarrow f(z) = (z-a)^n g_n(z) \neq 0. \quad \clubsuit$$

3. Obtain an alternative form of Cauchy's integral formula expressed by

$$f(z) = f(re^{i\phi}) = \frac{R^2 - r^2}{2\pi} \int_0^{2\pi} \frac{f(re^{i\theta})}{R^2 - 2rR \cos(\theta - \phi) + r^2} d\theta$$

that is valid for  $|z| < R$  if  $f(z)$  is analytic for  $|z| \leq R$ . This is called **Poisson's integral formula**.

**Solution:** Consider the function

$$g(\zeta) = \frac{z^*}{R^2 - z^*\zeta} f(\zeta),$$

which is analytic for  $|\zeta| \leq R$ . Hence, for the contour  $C : |\zeta| = R$ , we have  $\oint_C g(\zeta) d\zeta = 0$ . Furthermore, Cauchy's integral formula tells us that  $\oint_C f(\zeta)/(\zeta - z) d\zeta = 0$ . From these two results, we obtain

$$\frac{1}{2\pi i} \oint_C \left( \frac{1}{\zeta - z} + \frac{z^*}{R^2 - z^*\zeta} \right) f(\zeta) d\zeta = \frac{R^2 - |z|^2}{2\pi i} \oint_C \frac{f(\zeta)}{(\zeta - z)(R^2 - z^*\zeta)} d\zeta = 0. \quad (7.44)$$

Setting  $z = re^{i\phi}$  and  $\zeta = Re^{i\theta}$ , we have

$$\begin{aligned} (\zeta - z)(R^2 - z^*\zeta) &= (Re^{i\theta} - re^{i\phi})(R^2 - re^{-i\phi}Re^{i\theta}) \\ &= R^2 e^{i\theta} [R^2 - 2rR \cos(\theta - \phi) + r^2]. \end{aligned}$$

Substituting in (7.44), we arrive at the desired formula. ♣

## 7.4 Series Representations

### 7.4.1 Circle of Convergence

We now turn to a very important notion: series representations of complex analytic functions. To begin with, we note (without proof) that most of the definitions and theorems in connection with the convergence of series of *real* numbers and real functions presented in Chap. 2 and 3 can be applied to complex counterparts with little or no change. Here we give a basic theorem regarding the convergence property of infinite power series consisting of complex numbers.

#### ♠ Theorem:

If the power series

$$\sum_{n=0}^{\infty} a_n z^n \quad (7.45)$$

converges at  $z = z_0 \neq 0$ , then it converges absolutely at every point of  $|z| < |z_0|$  and, furthermore, it converges uniformly for  $|z| \leq \rho$  where  $0 < \rho < |z_0|$ .

*Proof* We first prove the statement regarding absolute convergence. From hypothesis, we see that the series  $\sum_{n=0}^{\infty} a_n z_0^n$  converges. We set

$$s_n = \sum_{k=0}^n a_k z_0^k,$$

to obtain

$$|s_n - s_{n-1}| = |a_n z_0^n| \rightarrow 0 \quad (n \rightarrow \infty).$$

Hence, there exists an integer  $M > 0$  that satisfies

$$|a_n z_0^n| \leq M \quad \text{for all } n,$$

which implies

$$\sum_{n=0}^{\infty} |a_n z^n| = \sum_{n=0}^{\infty} |a_n z_0^n| \left| \frac{z}{z_0} \right| \leq M \sum_{n=0}^{\infty} \left| \frac{z}{z_0} \right|^n.$$

Therefore, if  $|z| < |z_0|$ , the right-hand side converges so that the series (7.45) converges absolutely.

Next we consider uniform convergence. For every  $z$  satisfying the relation  $|z| \leq \rho < |z_0|$ , we have

$$\sum_{n=0}^{\infty} |a_n z^n| \leq M \sum_{n=0}^{\infty} \frac{\rho^n}{|z_0|^n},$$

since  $0 < \rho/|z_0| < 1$ . In view of the Weierstrass  $M$ -test, we conclude that the series (7.45) converges uniformly on the region of  $|z| \leq \rho$ . ♣

This theorem states that converging behavior of power series

$$\sum_{n=0}^{\infty} a_n z^n \tag{7.46}$$

can be classified into the following three types:

1. It converges at all  $z$ .
2. It converges (ordinary and thus absolutely) at  $|z| < R$ , but diverges at  $|z| > R$ , in which the real constant  $R$  depends on the feature of the series.
3. It diverges at all  $z$  except the origin.

This classification leads us to introduce the concept of **radius of convergence**  $R$  of the power series (7.46). For the above three cases, it becomes

1.  $R = 0$ ,   2.  $R$  itself,   3.  $R = \infty$ ,

respectively. The circle  $C$  with the radius  $R$  about the origin is called the **circle of convergence** associated with the series. Note that just on  $C$ , converging behavior of the corresponding series is inconclusive—it may or may not converge.

The following theorems provide us with a clue for finding the radius of convergence of a given power series.

♠ **Theorems:**

Given a power series  $\sum_{n=0}^{\infty} a_n z^n$ , its radius of convergence  $R$  equals

(i)  $R = \lim_{n \rightarrow \infty} \left| \frac{a_n}{a_{n+1}} \right|$ , if the limit exists;

(ii)  $R = \frac{1}{\limsup_{n \rightarrow \infty} \sqrt[n]{|a_n|}}.$

### 7.4.2 Singularity on the Radius of Convergence

Given a complex-valued power series, the convergence criterion based on the radius of convergence discussed in the previous subsection does not provide us with any information about the convergence property of the series just on the circle of convergence. We present below two important theorems regarding the latter point.

♠ **Theorem:**

If the power series  $\sum_{n=0}^{\infty} a_n z^n$  has a radius of convergence  $R$ , then it has at least one singularity on the circle  $|z| = R$ .

*Proof* Set

$$f(z) = \sum_{n=0}^{\infty} a_n z^n.$$

If  $f(z)$  were analytic at every point on the circle of convergence, then for each  $z$  with  $|z| = R$ , there would exist some maximal  $\varepsilon_{z_0}$  such that  $f(z)$  could be **continued analytically** to a circular region  $|z - z_0| < \varepsilon_{z_0}$  where  $z_0$  is located on the circle  $|z| = R$ . (See Sect. 8.3 for details of **analytic continuation**.) Here  $\varepsilon_{z_0}$  would depend on  $z_0$  and we define

$$\varepsilon \equiv \min_{|z_0|=R} \varepsilon_{z_0} > 0.$$

By performing continuations successfully for all possible  $z_0$ , we obtain a function  $g(z)$  that is analytic for  $|z| < R + \varepsilon$ . Clearly for  $|z| < R$ ,  $g$  must be identical to  $f$ . In addition,  $g$  must have a power series representation,

$$g(z) = \sum_{n=0}^{\infty} b_n z^n, \quad (7.47)$$

that is convergent for  $|z| < R + \varepsilon$ . Yet since for  $|z| < R$

$$g(z) = f(z) = \sum_{n=0}^{\infty} a_n z^n,$$

we conclude that

$$a_n \equiv b_n.$$

This implies that the radius of convergence of (7.47) would be  $R$ , which clearly gives us a contradiction. We thus conclude that  $f(z)$  has at least one singularity on the circle  $|z| = R$ . ♣

In general, it is difficult to determine when a function has a singularity at a particular point on the circle of convergence of its power series. The following theorem is one of the few results we have in this direction.

♠ **Theorem:**

Suppose that a power series  $\sum_{n=0}^{\infty} a_n z^n$  has a radius of convergence  $R < \infty$  and that  $a_n \geq 0$  for all  $n$ . Then the series has a singularity at  $z = R$  on the real axis.

*Proof* By the previous theorem, the function

$$f(z) = \sum_{n=0}^{\infty} a_n z^n$$

has a singularity at some point  $Re^{i\alpha}$ . If we consider the power series for  $f$  about a point  $\rho e^{i\alpha}$  with  $0 < \rho < R$ , we have

$$f(z) = \sum_{n=0}^{\infty} b_n (z - \rho e^{i\alpha})^n = \sum_{n=0}^{\infty} \frac{f^{(n)}(\rho e^{i\alpha})}{n!} (z - \rho e^{i\alpha})^n,$$

where the radius of convergence is  $R - \rho$ . (If it were larger, the power series would define an **analytic continuation** of  $f$  beyond  $Re^{i\alpha}$ .) Note, however, that for any nonnegative integer  $j$ , the derivative  $f^{(j)}$  reads

$$f^{(j)}(\rho e^{i\alpha}) = \sum_{n=j}^{\infty} n(n-1) \cdots (n-j+1) a_n (\rho e^{i\alpha})^{n-j}.$$

Since  $a_n \geq 0$ , we have

$$\left| f^{(j)}(\rho e^{i\alpha}) \right| \leq f^{(j)}(\rho).$$

This implies that the power series representation of  $f$  around,  $z = \rho$ , expressed by

$$f(z) = \sum_{n=0}^{\infty} \frac{f^{(n)}(\rho)}{n!} (z - \rho)^n$$

must have a radius of convergence  $R - \rho$ . On the other hand, if  $f$  were analytic at  $z = R$ , the above power series would converge on a disc of radius greater than  $R - \rho$ . Therefore,  $f$  is singular at  $z = R$ . ♣

### 7.4.3 Taylor Series

Below is the one of the main theorems of this section, which states that any analytic function can be expanded into a power series around its analytic point.

♠ **Taylor series expansion:**

If  $f(z)$  is analytic within and on the circle  $C$  of radius  $r$  around  $z = a$ , then there exists a unique and uniformly convergent series in powers of  $(z - a)$ ,

$$f(z) = \sum_{k=0}^{\infty} c_k (z - a)^k \quad (|z - a| \leq r), \quad (7.48)$$

with

$$c_k = \frac{f^{(k)}(a)}{k!} = \frac{1}{2\pi i} \oint_C \frac{f(\zeta)}{(\zeta - a)^{k+1}} d\zeta.$$

The largest circle  $C$  for which the power series (7.48) converges is called the **circle of convergence** of the power series and its radius is called the **radius of convergence**.

*Proof* Let  $f(z)$  be analytic within and on a closed contour  $C$ . From Cauchy's integral formula, we have

$$f(a + h) = \frac{1}{2\pi i} \oint_C \frac{f(z)}{z - a - h} dz, \quad (7.49)$$

where  $a$  is inside a contour  $C$ . The contour is taken to be a circle about  $a$ , inasmuch as the region of convergence of the resulting series is circular. We employ the identity

$$\left[ 1 + \frac{h}{z - a} + \frac{h^2}{(z - a)^2} + \cdots + \frac{h^{N-1}}{(z - a)^{N-1}} \right] \left( \frac{z - a - h}{z - a} \right) = 1 - \frac{h^N}{(z - a)^N}$$

to obtain the exact expression

$$\frac{1}{z - a - h} = \sum_{n=0}^{N-1} \left[ \frac{h^n}{(z - a)^{n+1}} \right] + \frac{h^N}{(z - a - h)(z - a)^N}.$$

Substituting this into (7.49), we have

$$f(a + h) = \sum_{n=0}^{N-1} \frac{h^n}{2\pi i} \oint_C \frac{f(z)}{(z - a)^{n+1}} dz + \frac{h^N}{2\pi i} \oint_C \frac{f(z)}{(z - a)^N (z - a - h)} dz. \quad (7.50)$$

Since the first integral can be replaced by the  $n$ th derivative of  $f$  at  $z = a$ , we have

$$f(a+h) = \sum_{n=0}^{N-1} \frac{h^n}{n!} f^{(n)}(a) + R_N, \quad (7.51)$$

where,  $R_N$  is the second term on the right-hand side of (7.50). It follows from (7.51) that if  $\lim_{N \rightarrow \infty} R_N = 0$ , the Taylor series expansion of  $f(z)$  around  $z = a$  is obtained successfully. This is indeed the case. As  $f(z)$  is analytic within and on the contour  $C$ , the absolute value of  $R_N$  is bounded as

$$|R_N| = \left| \frac{h^N}{2\pi i} \oint_C \frac{f(z)}{(z-a)^N(z-a-h)} dz \right| \leq \frac{|h|^N M r}{r^N(r-|h|)}, \quad (7.52)$$

where  $r$  is the radius of the circle and  $M$  is the maximum value of  $|f|$  on the contour. Within the radius  $r$ ,  $|h| < r$  so that

$$\lim_{N \rightarrow \infty} R_N = 0.$$

Hence, we have

$$f(a+h) = \sum_{n=0}^{\infty} \frac{h^n}{n!} f^{(n)}(a), \quad (7.53)$$

which holds at any point  $z = a + h$  within the circle of radius  $r$ . ♣

We note that the series (7.53) converges for large  $h$  as far as  $|h| < r_c$ , since  $R_N$  vanishes as  $N \rightarrow \infty$  for any value of  $|h|$  smaller than  $r_c$ . Furthermore, as the inequality (7.52) holds whenever  $f(z)$  is analytic within and on the circle of radius  $r_c$ , the radius of convergence,  $r$ , can extend up to the singularity is nearest neighbor to  $z = a$ . When the extending circle goes beyond the nearest singular point, the inequality becomes invalid so that the Taylor series expansion fails.

#### 7.4.4 Apparent Paradoxes

We have seen that the radius of convergence is determined by the distance to the nearest singularity. Interestingly, this explains some apparent paradoxes that which occur if we restrict our attention only to values of the series along the real axis of  $z$ .

A familiar example is the Taylor expansion of  $f(z) = 1/(1-z)$  around the origin:

$$\frac{1}{1-z} = 1 + z + z^2 + \cdots. \quad (7.54)$$

Obviously, both sides of (7.54) “blow up” at  $z = 1$ . At  $z = -1$ , on the other hand, the right-hand side diverges, whereas the left-hand side has a finite value of  $1/2$ . Notably, this apparent paradox occurs at all points represented

by  $z = e^{i\phi}$ , i.e., at any point on a unit circle surrounding the origin. The reason for this is clear from the point of view of the radius of convergence. (We leave it to the reader.)

Another example is

$$f(z) = e^{-1/z^2}.$$

Observe that  $f^{(n)}(0) = 0$  for any  $n = 0, 1, \dots$ , so if one puts this result blindly into the Taylor formula around  $z = 0$ , one obtains apparent nonsense as  $e^{-1/z^2} = 0$ . The point here is that  $z = 0$  is a singularity, where the Taylor series expansion is prohibited.

These two examples suggest the importance of realizing the difference between the series representing a function and “the function itself.” A power series, such as a Taylor series, has only a limited range of representation characterized by the radius of convergence. Beyond this range, the power series is unable to represent the function. For example, the function considered in (7.54),

$$f(z) = \frac{1}{1-z}, \quad (7.55)$$

*exists* and is analytic everywhere except at  $z = 1$ , but its power series around  $z = 0$ , given by

$$1 + z + z^2 + \dots$$

exists and represents  $f$  *only* within the unit circle centered at the origin (i.e.,  $|z| < 1$ ). The region in which a power series reproduces its original function is dependent on the explicit form of the series expansion. In fact, an alternative series expansion of (7.55) around  $z = 3$  is given by

$$-\frac{1}{2} + \frac{1}{4}(z-3) - \frac{1}{8}(z-3)^2 + \dots,$$

which exists and represents (7.55) only within the circle of radius 2 centered at  $z = 3$ . We thus conclude that power series (including Taylor’s, Laurent’s, and others) are not regarded as pieces of a versatile mold by means of which one can cast a copy of the function. Each piece of the mold can reproduce the behavior of  $f$  only within the region where the series converges, but gives no indication of the shape of  $f$  beyond its range.

### 7.4.5 Laurent Series

When expanding a function  $f(z)$  around its singular point  $z = a$ , Taylor’s expansion is obviously not suitable but we can obtain an alternative expansion that is valid for a singular point. The latter kind of expansion is called a **Laurent series expansion**. Laurent series enter quite often in mathematical analyses of physical problems, where functions to be considered have a finite number of singularities.

♠ **Laurent series expansions:**

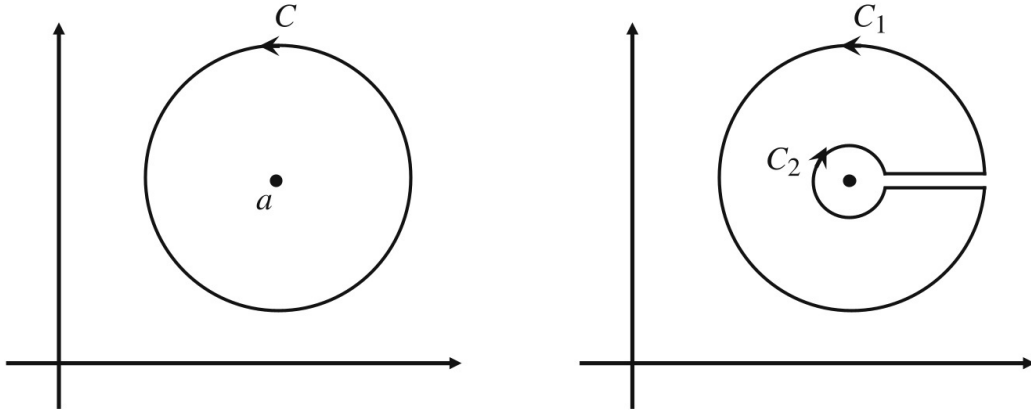
Let  $f(z)$  be analytic within and on a closed contour  $C$  except at a point  $z = a$  enclosed by  $C$ . Then,  $f(z)$  can be expanded around  $z = a$  as

$$f(z) = \sum_{n=-\infty}^{\infty} c_n (z - a)^n, \quad (7.56)$$

with the definition

$$c_n = \frac{1}{2\pi i} \oint_C \frac{f(\zeta)}{(\zeta - a)^{n+1}} d\zeta. \quad (7.57)$$

The series (7.56) with the constants (7.57) is called the Laurent series expansion of  $f(z)$ .



**Fig. 7.12.** Conversion of a closed contour  $C$  into  $C_1 + C_2$  so as not to involve the singularity of  $f(z)$  at  $z = a$  in it

*Proof* The trick to deriving a Laurent series expansion is to use the contour  $C_1 + C_2$  illustrated in Fig. 7.12 such that its interior does not contain the singular point of  $f(z)$  at  $z = a$  (i.e.,  $f$  is analytic within and on the contour). As is indicated, the original contour  $C$  can be reduced to two circular contours  $C_1$  and  $C_2$  encircling  $z = a$  counterclockwise and clockwise, respectively. Applying Cauchy's theorem, we have

$$\begin{aligned} f(a + h) &= \frac{1}{2\pi i} \oint_C \frac{f(z)}{z - a - h} dz \\ &= \frac{1}{2\pi i} \oint_{C_1} \frac{f(z)}{z - a - h} dz - \frac{1}{2\pi i} \oint_{C_2} \frac{f(z)}{z - a - h} dz. \end{aligned} \quad (7.58)$$

Note that  $|z - a| > |h|$  on the contour  $C_1$  and  $|z - a| < |h|$  on  $C_2$ . We thus have

$$\frac{1}{z - a - h} = \frac{1}{z - a} \cdot \frac{1}{1 - \frac{h}{z - a}} = \frac{1}{z - a} \sum_{n=0}^{\infty} \left( \frac{h}{z - a} \right)^n \quad \text{on } C_1 \quad (7.59)$$

and

$$\frac{1}{z - a - h} = \frac{1}{h} \cdot \frac{1}{\frac{z - a}{h} - 1} = \frac{-1}{h} \sum_{n=0}^{\infty} \left( \frac{z - a}{h} \right)^n \quad \text{on } C_2. \quad (7.60)$$

The substitution of these two expressions into (7.58) yields

$$f(a + h) = \frac{1}{2\pi i} \left[ \oint_{C_1} \sum_{n=0}^{\infty} \frac{h^n}{(z - a)^{n+1}} f(z) dz + \oint_{C_2} \sum_{n=1}^{\infty} \frac{(z - a)^{n-1}}{h^n} f(z) dz \right]. \quad (7.61)$$

The order of integration and summation within the square brackets can be reversed since the infinite series involved in the integrals converge. Eventually, we obtain

$$f(a + h) = \sum_{n=-\infty}^{\infty} c_n h^n; \quad c_n = \frac{1}{2\pi i} \oint_C \frac{f(z)}{(z - a)^{n+1}} dz. \quad (7.62)$$

Here, the contour for the coefficients  $c_n$  should be  $C_1$  in the positive direction for  $n \geq 0$  and  $C_2$  in the negative direction for  $n < 0$ . The series (7.62) is what we call the Laurent series expansion of  $f(z)$  around the singular point  $z = a$ . Note that  $C_1$  can be taken as the contour for all values of  $n$  with the reverse direction for negative  $n$ 's. This is because the integrand is analytic in the region between  $C_1$  and  $C_2$ , which allows us to expand the size of the contour  $C_2$  until it coincides with the larger contour  $C_1$ . ♣

#### 7.4.6 Regular and Principal Parts

An important property of Laurent series is the series resolution. To see this, we rewrite (7.62) as follows:

$$f(a + h) = \sum_{n=0}^{\infty} c_n h^n + \sum_{n=1}^{\infty} c_{-n} h^{-n}. \quad (7.63)$$

The first term in (7.63) converges everywhere within the outer circle of convergences, whereas the second term converges anywhere outside the inner circle. This means that the Laurent series expansion resolves the original function  $f(z)$  into two parts: one that is analytic within the outer circle of convergence, and the other that is analytic outside the inner circle of convergence. Obviously, each part is analytic over different portions of the complex plane.

The part of the Laurent series consisting of positive powers of  $h$  is called the **regular part**. The other part, consisting of negative powers, is called the **principal part**. Either part (or both) may terminate at a finite degree of the sum or be identically zero. Particularly when the principal part is identically zero, then  $f(z)$  is analytic at  $z = a$ , and the Laurent series is identical with the Taylor series.

**Remark.** At first glance, the regular part exhibited in (7.63) resembles the Taylor series. However, this is not the case; the  $n$ th coefficient cannot generally be associated with  $f^{(n)}(a)$  because the latter may not exist. In most applications,  $f(z)$  is not analytic at  $z = a$ .

### 7.4.7 Uniqueness of Laurent Series

Taylor and Laurent series allow us to express an analytic function as a power series. For a Taylor series of  $f(z)$ , the expansion is routine because the coefficient of its  $n$  term is simply  $f^{(n)}(z_0)/n!$ , where  $z_0$  is the center of the circle of convergence. In contrast, for the case of a Laurent series expansion, the  $n$ th coefficient is not (in general) easy to evaluate. It can usually be found by inspection and certain manipulations of other known series, but if we use such an intuitive approach to determine the coefficients, we cannot be sure that the result we obtain is correct. The following theorem addresses this issue.

♠ **Theorem:**

If the series

$$\sum_{n=-\infty}^{\infty} a_n(z - z_0)^n \quad (7.64)$$

converges to  $f(z)$  at all points in some annular region around  $z_0$ , then it is the unique Laurent series expansion of  $f(z)$  in that region.

*Proof* Multiply both sides of (7.64) by

$$\frac{1}{2\pi i(z - z_0)^{k+1}},$$

integrate the result along a contour  $C$  in the annular region, and use the easily verifiable fact that

$$\frac{1}{2\pi i} \oint_C \frac{dz}{(z - z_0)^{k-n+1}} = \delta_{kn}$$

to obtain

$$\frac{1}{2\pi i} \oint_C \frac{f(z)}{(z - z_0)^{k+1}} = a_k.$$

Thus, the coefficient  $a_k$  in the power series (7.64) is precisely the coefficient in the Laurent series of  $f(z)$  given in (7.57), and the two must be identical. ♣

*Remark.* A Laurent series is unique only for a specified annulus. In general, a function  $f(z)$  can possess two or more entirely different Laurent series about a given point, valid for different (nonoverlapping) regions; For instance,

$$f(z) = \frac{1}{z(1-z)} = \begin{cases} \frac{1}{z} + 1 + z + z^2 + \cdots, & 0 < |z| < 1, \\ -\frac{1}{z^2} - \frac{1}{z^3} - \frac{1}{z^4} - \cdots, & 1 < |z| < \infty. \end{cases}$$

### 7.4.8 Techniques for Laurent Expansion

The following examples illustrate several useful techniques for the construction of Taylor and Laurent series.

#### (a) Use of geometric series

Suppose that a function

$$f(z) = \frac{1}{z-a} \quad (7.65)$$

fails to be analytic at  $z = a$ . We would like to obtain the Laurent series of  $f(z)$  around  $z = a$ . First we note that for  $|z| < |a|$ ,  $f(z)$  reads

$$\frac{1}{z-a} = -\frac{1}{a} \frac{1}{1-(z/a)} = -\frac{1}{a} \sum_{n=0}^{\infty} \left(\frac{z}{a}\right)^n. \quad (7.66)$$

This is obviously the Taylor series expansion of  $f(z)$  around the point  $z = 0$ . That is, for  $|z| < |a|$ , the Laurent series of  $f(z)$  given in (7.65) becomes identical to its Taylor series. Nevertheless this is not the case for  $|z| > |a|$ , since its radius of convergence is  $R = |a|$ . Hence, we should also evaluate the Laurent series around  $z = a$  that is valid for  $|z| > |a|$ . In a similar manner as above, we obtain

$$\frac{1}{z-a} = \frac{1}{z} \sum_{n=0}^{\infty} \left(\frac{a}{z}\right)^n = \sum_{n=0}^{\infty} \frac{a^n}{z^{n+1}} \quad \text{for } |z| > |a|. \quad (7.67)$$

Expansions (7.66) and (7.67) both serve as the Laurent series expansions of  $f(z)$ , although the regions of convergence are different from one another.

*Remark.* The function  $f(z)$  given in (7.65) can be expanded by this method about any point  $z = b$ ; Indeed, write

$$f(z) = \frac{1}{z-a} = \frac{1}{(z-b) - (a-b)} \quad (b \neq a).$$

Then, either

$$f(z) = -\frac{1}{a-b} \sum_{n=0}^{\infty} \frac{(z-b)^n}{(a-b)^n} \quad (|z-b| < |a-b|)$$

or

$$f(z) = \sum_{n=0}^{\infty} \frac{(a-b)^n}{(z-b)^{n+1}} \quad (|z-b| > |a-b|).$$

### (b) Rational fraction decomposition

Next we assume a function

$$f(z) = \frac{1}{z^2 - (2+i)z + 2i}.$$

The roots of the denominator are  $z = i$  and  $z = 2$ , which are the only points at which  $f(z)$  fails to be analytic. Hence,  $f(z)$  has a Taylor series about  $z = 0$  that is valid for  $|z| < 1$  and two Laurent series about  $z = 0$  that are valid for  $1 < |z| < 2$  and  $|z| > 2$ . To obtain them, we use the identities

$$z^2 - (2+i)z + 2i = (z-i)(z-2)$$

and

$$f(z) = \frac{1}{(z-i)(z-2)} = \frac{1}{2-i} \left( \frac{1}{z-2} - \frac{1}{z-i} \right).$$

When we want the Laurent series of  $f(z)$  around  $z = 0$  that is valid for  $1 < |z| < 2$ , it suffices to expand the function  $1/(z-2)$  in the Taylor series about  $z = 0$  [see (a) above] and then expand  $1/(z-i)$  in the Laurent series about  $z = 0$  that is valid for  $|z| > 1$ . (The latter series is also valid for  $1 < |z| < 2$ .) If these two series are subtracted, we obtain a series for  $f(z)$  that is valid for  $1 < |z| < 2$ , which is the desired Laurent series.

### (c) Differentiation

The method used in (b) fails for functions with a double root in the denominator such that

$$f(z) = \frac{1}{(z-1)^2}.$$

Among alternative methods, the simplest one is the differentiation

$$\frac{1}{(z-1)^2} = \frac{d}{dz} \left( \frac{1}{1-z} \right).$$

From the discussions regarding the earlier case **(a)**, the function  $1/(1-z)$  is seen to be represented by

$$\frac{1}{1-z} = \begin{cases} \sum_{n=0}^{\infty} z^n, & |z| < 1, \\ -\sum_{n=0}^{\infty} \frac{1}{z^{n+1}}, & |z| > 1. \end{cases}$$

Hence, term-by-term differentiations yield

$$\frac{1}{(z-1)^2} = \begin{cases} \sum_{n=0}^{\infty} (n+1)z^n, & |z| < 1, \\ -\sum_{n=0}^{\infty} (n+1)z^{-(n+2)}, & |z| > 1. \end{cases}$$

## Exercises

1. Let  $f(z)$  be an entire function. Employ the Taylor series expansion to show that the function defined by

$$g(z) = \begin{cases} \frac{f(z) - f(a)}{z - a}, & z \neq a, \\ f'(a), & z = a. \end{cases}$$

is also entire.

**Solution:** For  $z \neq a$ , we employ the Taylor series expansion of  $f(z)$  to obtain

$$g(z) = f'(a) + \frac{f''(a)}{2!}(z-a) + \frac{f^{(3)}(a)}{3!}(z-a)^2 + \cdots \quad (7.68)$$

By the definition of  $g$ , the representation (7.68) is valid at  $z = a$ . Hence,  $g$  is equal to an everywhere-convergent power series and is thus an entire function. ♣

2. If  $f$  is entire and if for some integer  $k \geq 0$  there exist positive constants  $A$  and  $B$  such that

$$|f(z)| \leq A + B|z|^k,$$

then  $f$  is a polynomial of degree  $k$  at most. Prove it

**Solution:** Note that the case  $k = 0$  is the original Liouville theorem. To prove the case of  $k > 0$ , we employ mathematical induction, and consider

$$g(z) = \begin{cases} \frac{f(z) - f(0)}{z}, & z \neq 0, \\ f'(0), & z = 0, \end{cases} \quad (7.69)$$

where  $f(z)$  is assumed to obey the conditions noted above. By Exercise 1,  $g$  is entire. In addition, by hypothesis on  $f$  we have

$$|g(z)| \leq C + D|z|^{k-1}.$$

Hence, by induction,  $g$  is a polynomial of degree  $k - 1$  at most, then  $f$  is polynomial of degree  $k$  at most owing to the definition (7.69). This completes the proof. ♣

**3.** Find the Laurent series of the multivalued logarithmic function given by

$$f(z) = \log(1 + z) = \log|1 + z| + i\arg(1 + z).$$

**Solution:** The **branch cut** (see Sect. 8.2.3) is set so as to extend from  $-\infty$  to  $-1$  along the real axis. Hence,  $\log(1 + z)$  is analytic within the circle  $|z| = 1$ . Since

$$\frac{d}{dz} \log(1 + z) = \frac{1}{1 + z},$$

we may expand

$$\frac{1}{1 + z} = 1 - z + z^2 - z^3 + \cdots = \sum_{n=0}^{\infty} (-1)^n z^n \quad (|z| < 1).$$

Then, term-by-term integration yields

$$\int^z \frac{d\xi}{1 + \xi} = z - \frac{z^2}{2} + \frac{z^3}{3} - \cdots + C \quad (|z| < 1),$$

where  $C$  is the constant of integration. Since  $\log 1 = 0$ , it follows that  $C = 0$  and

$$\log(1 + z) = z - \frac{z^2}{2} + \frac{z^3}{3} - \cdots = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{z^n}{n} \quad (|z| < 1).$$

Other branches of  $\log(1 + z)$  have the same series except for different values of the constant  $C$ . ♣

4. Find the power series representation of  $f(z)$  about  $z = 0$  that satisfies the differential equation

$$f'(z) + f(z) = 0 \quad \text{with } f(0) = 1. \quad (7.70)$$

**Solution:** Let  $f(z) = 1 + \sum_{n=1}^{\infty} a_n z^n$ . Then we have  $f'(z) = \sum_{n=1}^{\infty} n a_n z^{n-1} = a_1 + \sum_{n=1}^{\infty} (n+1) a_{n+1} z^n$ . Substitute this into (7.70) to obtain

$$1 + a_1 = 0 \quad \text{and} \quad a_n + (n+1) a_{n+1} = 0 \quad \text{for } n \geq 1.$$

The latter result yields

$$a_n = (-1) \frac{1}{n} a_{n-1} = (-1)^2 \frac{1}{n(n-1)} a_{n-2} = \cdots = (-1)^n \frac{1}{n!} a_1.$$

Hence, we have  $a_n = (-1)^n / n!$ , so that

$$f(z) = 1 + \sum_{n=1}^{\infty} \frac{(-1)^n}{n!} z^n = e^{-z}. \quad \clubsuit$$

5. Let  $f(z) = \sum_{n=0}^{\infty} c_n (z-a)^n$  be analytic for  $|z-a| < R$ . Prove that

$$\frac{1}{2\pi} \int_0^{2\pi} |f(a + re^{i\theta})|^2 d\theta = \sum_{n=0}^{\infty} |c_n|^2 r^{2n} \quad \text{for any } r < R.$$

Then show that

$$\sum_{n=0}^{\infty} |c_n|^2 r^{2n} \leq M(r)^2, \quad (7.71)$$

in which  $M(r) = \max_{|z-a|=r} |f(z)|$ . The result (7.71) is called **Gutzmer's theorem**.

**Solution:** From assumption, it follows that

$$|f(z)|^2 = \left[ \sum_{n=0}^{\infty} c_n (re^{i\theta})^n \right] \left[ \sum_{m=0}^{\infty} c_m^* (re^{-i\theta})^m \right] = \sum_{n,m=0}^{\infty} c_n c_m^* r^{m+n} e^{i(n-m)\theta}.$$

This infinite series converges uniformly on the circle  $|z-a| = r < R$ , which allows us to interchange the order of integration and summation as expressed by

$$\int_0^{2\pi} |f(a + re^{i\theta})|^2 d\theta = \sum_{n,m=0}^{\infty} c_n c_m^* r^{m+n} \int_0^{2\pi} e^{i(n-m)\theta} d\theta.$$

The right-hand side vanishes when  $n \neq m$  since the integral equals zero. Hence, we have

$$\int_0^{2\pi} |f(a + re^{i\theta})|^2 d\theta = \sum_{n=0}^{\infty} |c_n|^2 r^{2n} \times 2\pi,$$

which is equivalent to the desired equation. Furthermore, since  $|f(a + re^{i\theta})| \leq M(r)$ , we have

$$\sum_{n=0}^{\infty} |c_n|^2 r^{2n} = \frac{1}{2\pi} \int_0^{2\pi} |f(a + re^{i\theta})|^2 d\theta \leq \frac{1}{2\pi} \int_0^{2\pi} M(r)^2 d\theta = M(r)^2. \quad \clubsuit$$

## 7.5 Applications in Physics and Engineering

### 7.5.1 Fluid Dynamics

This section demonstrates the effectiveness of using complex function theory for analyzing fluid dynamics in a two-dimensional plane. The primary aim is to derive the **Kutta–Joukowski theorem** (see Sect. 7.5.2), which describes the **lift force** exerted on a solid material placed in a uniform flow. Before proceeding, we introduce terminologies and several basic concepts that pertain to fluid dynamics.

The fundamental quantities that characterize a two-dimensional fluid flow are velocity  $\mathbf{v} = u\mathbf{e}_x + v\mathbf{e}_y$  and vorticity  $\boldsymbol{\omega} = \nabla \times \mathbf{v}$ , both of which are vector-valued functions of the position  $\mathbf{r}$ . Here, we restrict our attention to the case of an **irrotational** ( $\boldsymbol{\omega} = \mathbf{0}$ ) and **incompressible** ( $\nabla \cdot \mathbf{v} = 0$ ) fluid. The assumption  $\boldsymbol{\omega} = \nabla \times \mathbf{v}$  allows us to define an appropriate function  $\Phi(x, y)$  such that

$$\mathbf{v} = \nabla \Phi, \tag{7.72}$$

since  $\nabla \times (\nabla f) = \mathbf{0}$  for any analytic function  $f(x, y)$  in the  $x$ - $y$  plane. The function  $\Phi(x, y)$  defined by (7.72) is called the **velocity potential**. Further, our assumption of  $\nabla \cdot \mathbf{v} = 0$  implies that

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0,$$

which in turn suggests the presence of an analytic function  $\Psi(x, y)$  defined by

$$u \equiv \frac{\partial \Psi}{\partial y}, \quad v \equiv -\frac{\partial \Psi}{\partial x} \tag{7.73}$$

that satisfies the two-dimensional Laplace equation  $\nabla^2 \Psi = 0$ . Such a function  $\Psi(x, y)$  is called a **stream function**.

*Remark.* The name *stream function* originates from the fact that the curves of  $\Psi(x, y) = \text{const.}$  in the  $x$ - $y$  plane represent streamline flow. This is shown by noting that if  $d\Psi = 0$ , we have

$$d\Psi = \frac{\partial\Psi}{\partial x}dx + \frac{\partial\Psi}{\partial y}dy = -vdx + udy = 0,$$

so that  $dx/u = dy/v$ , which implies that  $d\mathbf{r}$  is parallel to  $\mathbf{v}$ .

From (7.72) to (7.73), it follows that the components of the velocity  $\mathbf{v}$  are expressed as

$$u = \frac{\partial\Phi}{\partial x} = \frac{\partial\Psi}{\partial y}, \quad v = \frac{\partial\Phi}{\partial y} = -\frac{\partial\Psi}{\partial x}.$$

This allows us to introduce the concept of a **complex velocity potential**  $f(z)$  in the complex plane:

$$f(z) = \Phi(z) + i\Psi(z) \quad \text{with } z = x + iy. \quad (7.74)$$

Note that since  $f(z)$  is analytic,

$$\frac{\partial f}{\partial x} = \frac{df}{dz} = u - iv = |\mathbf{v}|e^{-i\theta},$$

i.e., the absolute value of the derivative  $|df/dz|$  gives the magnitude of the velocity  $|\mathbf{v}|$ . Furthermore, the contour integral of  $f(z)$  has important physical implications. Given a closed contour  $C$  placed on a two-dimensional flow, we have

$$\oint_C df = \Gamma(C) + iQ(C),$$

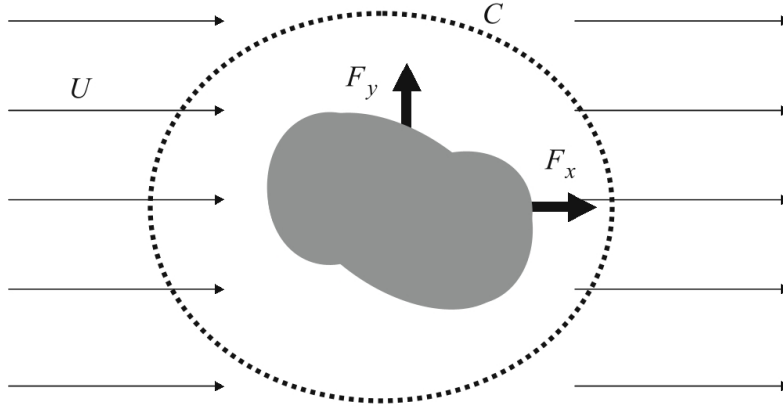
where

$$\begin{aligned} \Gamma(C) &= \oint_C d\Phi = \oint_C (u dx + v dy) = \oint_C \mathbf{v} \cdot d\mathbf{r}, \\ Q(C) &= \oint_C d\Psi = \oint_C (u dy - v dx) = \oint_C |\mathbf{v} \times d\mathbf{r}|. \end{aligned}$$

Hence, the integrals  $\Gamma(C)$  and  $Q(C)$  represent the **circulation** (or **rotation**) and the **fluid flow**, respectively.

### 7.5.2 Kutta–Joukowski Theorem

We are now ready to study the Kutta–Joukowski theorem, which describes the lift force in a two-dimensional flow. The lift force is a component of the fluid dynamic force that is perpendicular to the flow direction. It is the lift force that makes it possible for airplanes, helicopters, sail boats, etc. to move against the gravitational force or water currents.



**Fig. 7.13.** Spatial configuration of material placed into a two-dimensional uniform flow with speed  $U$  showing the components  $F_x, F_y$  of the flow-induced force  $\mathbf{F}$  acting on the material

♠ **Kutta–Joukowski theorem:**

The lift force  $F_y$  that acts on a material placed in a uniform flow  $U$  in the  $x$ -direction is given by

$$F_y = -\rho U \Gamma(C), \quad (7.75)$$

where  $\rho$  and  $\Gamma(C)$  are the mass density and the circulation of the fluid, respectively, within a closed contour  $C$  surrounding the material (see Fig. 7.13).

The lift force is generated in accordance with **Bernoulli's theorem** and the **law of conservation of momentum**. Both of these principles are used to explain the mechanism responsible for the occurrence of the lift force in a uniform flow, which is given by the **Blasius formula** (see 7.5.3):

$$\mathbf{F} = \frac{i\rho}{2} \oint_C w^2 dz,$$

which plays a key role in the proof of the Kutta–Joukowski theorem, as shown below.

*Proof (of the Kutta–Joukowski theorem).* Assume a uniform flow oriented to the  $x$ -axis. Then the function  $w = df/dz$  is analytic and satisfies the relation

$$\lim_{z \rightarrow \infty} w = U = \text{const.}$$

Hence,  $w$  can be expanded at points sufficiently far from the origin:

$$w = \frac{df}{dz} = U + \frac{k_0}{z} - \frac{c_1}{z^2} - \frac{2c_2}{z^3} - \cdots \quad (z \rightarrow \infty), \quad (7.76)$$

which implies that

$$f = Uz + k_0 \log z + c_0 + \frac{c_1}{z} + \frac{c_2}{z^2} + \cdots \quad (z \rightarrow \infty) \quad (7.77)$$

and

$$w^2 = U^2 + \frac{2Uk_0}{z} + (k_0^2 - 2Uc_1) \frac{1}{z^2} + \cdots \quad (z \rightarrow \infty). \quad (7.78)$$

From (7.77) we have

$$\oint_C f dz = 2\pi i k_0 = \Gamma(C) + iQ(C), \quad (7.79)$$

and substituting (7.78) into the **Blasius formula** expressed by  $\mathbf{F} = (i\rho/2) \oint_C w^2 dz$ , we obtain

$$\mathbf{F} = F_x + iF_y = \frac{i\rho}{2} \cdot 2\pi i \cdot 2Uk_0 = -2\pi\rho U k_0. \quad (7.80)$$

Combining (7.79) and (7.80) yields

$$F_x + iF_y = \rho U(-Q + i\Gamma),$$

i.e.,

$$F_x = -\rho U Q, \quad F_y = -\rho U \Gamma. \quad (7.81)$$

of the two results above, it is the second one regarding  $F_y$  that states the theorem. ♣

**Remark.** The first equation in (7.81) indicates that  $F_x = 0$  if  $Q = 0$ ; i.e. no force in the direction of the stream is relevant to a material inside the closed contour  $C$  if no source is located interior to  $C$ . This is precisely the case for an ideal flow without any viscosity.

### 7.5.3 Blasius Formula

We conclude this section by explaining the **Blasius formula**, which is important for the proof of the Kutta–Joukowski theorem discussed above. Consider a two-dimensional flow of irrotational and incompressible fluid and assume that a solid material is placed inside a closed contour  $C$  encircling a portion of the fluid. Apparently, a force  $\mathbf{F}$  from the flow is exerted on the material. Hence, the **law of the conservation of momentum** within the contour  $C$  is written as

$$\mathbf{F} + \oint_C d\mathbf{G} = 0,$$

where  $d\mathbf{G}$  represents the sum of momentums that pass through a line element  $ds$  of the closed contour  $C$  per unit time. It is given by

$$d\mathbf{G} = p\mathbf{n}ds + \rho\mathbf{v}v_nds, \quad (7.82)$$

where  $p$  is the fluid pressure,  $\mathbf{n}$  is a basis vector normal to the contour  $C$ ,  $\rho$  is the density of the fluid, and  $v_n = \mathbf{v} \cdot \mathbf{n}$ . The first and second terms on the right-hand side of (7.82) represent the impulse transmitted to the interior of  $C$  through  $ds$  and the volume of fluid passing through  $ds$ , respectively. Using the stream potential  $\Psi$ , we rewrite as (7.82)

$$d\mathbf{G} = p\mathbf{n}ds + \rho v d\Psi, \quad (7.83)$$

since  $d\Psi = v_n ds$ .

In order to obtain the complex-number representation of (7.83), we denote by  $dz$  an infinitesimal vector having length  $ds$  and a direction normal to  $\mathbf{n}$ . We then have

$$dz = i(n_x + in_y)ds.$$

when we apply this relation to (7.82), the quantity  $d\mathbf{G}$  is expressed as

$$dG_x + idG_y = -ipdz + \rho \frac{df^*}{dz^*} \cdot \frac{df - df^*}{2i}, \quad (7.84)$$

where we consider  $d\Psi$  to be the imaginary part of  $df$ . The pressure  $p$  is known to correlate with  $f$  via **Bernoulli's theorem**, which is expressed by

$$p = p_0 - \frac{\rho}{2} \left| \frac{df}{dz} \right|^2 = p_0 - \frac{\rho}{2} \frac{df}{dz} \frac{df^*}{dz^*}, \quad (7.85)$$

where  $p_0$  is the pressure at a position far from the material (i.e.,  $z \rightarrow \infty$ ). It then follows from (7.84) to (7.85) that

$$dG_x + idG_y = -ip_0dz + \frac{i\rho}{2} \frac{df}{dz} \frac{df^*}{dz^*} dz - \frac{i\rho}{2} \frac{df^*}{dz^*} \left( \frac{df}{dz} dz - \frac{df^*}{dz^*} dz^* \right) \quad (7.86)$$

$$= -ip_0dz - \frac{i\rho}{2} \left( \frac{df}{dz} \right)^2 dz. \quad (7.87)$$

Since  $\oint_C dz = 0$ , we finally obtain

$$\mathbf{F} = F_x + iF_y = \frac{i\rho}{2} \oint_C \left( \frac{df}{dz} \right)^2 dz, \quad (7.88)$$

which is known as the **Blasius formula**.

## Singularity and Continuation

---

**Abstract** We devote the first half of this chapter to the essential properties and classification of singularities, which are nonanalytic points in a complex plane. We then describe analytic continuation, which is a most important concept from a theoretical as well as an applied point of view. Through analytic continuations, we observe the interesting fact that the functional form of a complex function may undergo various changes depending on the defining region in the complex plane.

### 8.1 Singularity

#### 8.1.1 Isolated Singularities

A **singularity** of a complex function  $f(z)$  is any point where it is not analytic. In particular, the point  $z = a$  is called an **isolated singularity** if and only if  $f(z)$  is analytic in some neighborhood but not at  $z = a$ . Most singularities we have encountered so far in this text were isolated singularities. However, we will see later that there are singularities that are not isolated.

When  $z = a$  is an isolated singularity of  $f(z)$ , it is classified as follows:

1. A **removable singularity** if and only if  $f(z)$  is finite throughout a neighborhood of  $z = a$ , except possibly at  $z = a$  itself.
2. A **pole of order  $m$**  ( $m = 1, 2, \dots$ ) if and only if  $(z - a)^m f(z)$  but not  $(z - a)^{m-1} f(z)$  is analytic at  $z = a$ . In this case,  $\lim_{z \rightarrow a} |f(z)| = \infty$  no matter how  $z$  approaches  $z = a$ .
3. An **essential singularity** if and only if the Laurent series of  $f(z)$  around  $z = a$  has an infinite number of terms involving negative powers of  $(z - a)$ .

**Remark.** There is an alternative definition of a pole: the point  $z = a$  is a pole of  $m$ th order of  $f(z)$  if and only if  $1/f(z)$  is analytic and has a **zero of order  $m$**  at  $z = a$ .

The three types of isolated singularities described above can be distinguished by the degree of expansion of the Laurent series of  $f(z)$  being considered. Let  $f(z)$  have an isolated singularity at  $z = a$ . Then there is a real number  $\delta > 0$  such that  $f(z)$  is analytic for  $0 < |z - a| < \delta$  but not for  $z = a$ , which means that  $f(z)$  can be represented by the Laurent series

$$f(z) = \sum_{n=0}^{\infty} c_n (z - a)^n + \sum_{n=0}^M c_{-n} \frac{1}{(z - a)^n}. \quad (8.1)$$

Thus, it suffices to examine the expansion degree  $M$  of the principal part, the second sum in (8.1), in order to determine the type of the isolated singularity  $z = a$ .

### Case 1. Removable singularities ( $M = 0$ )

In this case, the principal part is absent so that the Laurent series around  $z = a$  reads

$$f(z) = c_0 + c_1(z - a) + c_2(z - a)^2 + \cdots \quad (z \neq a).$$

Observe that  $\lim_{z \rightarrow a} f(z) = c_0$  as is consistent with statement 1 above, which says that  $f(z)$  is finite in a neighborhood of  $z = a$ . This kind of singularity can be eliminated by redefining  $f(a)$  as  $c_0$ , which is why we call it *removable*.

*Examples* Consider the function

$$f(z) = \frac{\sin z}{z}. \quad (8.2)$$

This yields  $\lim_{z \rightarrow 0} f(z) = 1$ , but the value of  $f(0)$  is not defined. Hence,  $z = 0$  is a removable singularity of (8.2). In a similar sense, the functions

$$e^{\sin z/z} \quad \text{and} \quad \frac{1}{z} - \frac{1}{\tan z}$$

are regarded as analytic at  $z = 0$ , since this point is the removable singularity for each.

### Case 2. Isolated poles ( $M$ is finite)

The second type of isolated singularity, for which the principal part reads

$$\sum_{n=1}^M c_{-n} h^{-n} \quad (c_{-M} \neq 0, M \geq 1),$$

is called a **pole** of order  $M$ . Order  $M$  is the minimum of the integer that makes the quantity

$$\lim_{z \rightarrow z_0} (z - z_0)^M f(z)$$

a finite, nonzero complex number.

*Examples* 1. The function  $f(z) = 1/\sin z$  has Laurent series valid for  $0 < |z| < \pi$ ;

$$\frac{1}{\sin z} = \frac{1}{z} + \frac{z}{6} + \frac{7}{360}z^3 + \frac{31}{15120}z^5 + \cdots,$$

from which it follows that it has a simple pole at the origin.

2. The function  $f(z) = 1/z$  has a simple pole at  $z = 0$ , which is easily seen by noting that  $\lim_{z \rightarrow 0} zf(z) = 1$ .

### Case 3: Essential singularities ( $M = \infty$ )

The third type of isolated singularity, **essential singularity**, gives rise to an infinite principal part.

*Examples* The function  $f(z) = e^{1/z}$  has the Laurent series

$$e^{1/z} = 1 + \frac{1}{z} + \frac{1}{2!z^2} + \frac{1}{3!z^3} + \cdots,$$

which is valid for  $|z| > 0$ . Since the principal part is infinite, the function has essential singularity at  $z = 0$ .

*Remark.* An infinite principal part in the Laurent series implies essential singularity *only when* the series is valid for all points in a neighborhood  $|z - a| < \varepsilon$  except  $z = a$ . For example, the series

$$f(z) = \frac{1}{(z-1)^2} + \frac{1}{(z-1)^3} + \frac{1}{(z-1)^4} + \cdots$$

does not mean that  $z = 1$  is an essential singularity of  $f(z)$ , since the series converges only if  $|z-1| > 1$ . It actually represents the function  $f(z) = 1/(z^2 - 3z + 2)$  in the annulus  $1 < |z-1| < R$ , which evidently has a simple pole at  $z = 1$ .

#### 8.1.2 Nonisolated Singularities

As noted earlier, there are other kinds of singular points that are neither poles nor essential singularities. For example, neither  $\sqrt{z}$  nor  $\log z$  can be expanded near  $z = 0$  in Laurent series; both of them are discontinuous along an entire line (say, the negative real axis) so that the singular point  $z = 0$  is not isolated. Singularities of this kind, called **branch points**, are discussed in the next subsection.

Another type of singular behavior of an analytic function occurs when it possesses an infinite number of isolated singularities converging to some limit point. Consider, for instance,

$$f(z) = \frac{1}{\sin(1/z)}.$$

The denominator has simple zeros whenever

$$z = \frac{1}{n\pi} \quad (n = \pm 1, \pm 2, \dots).$$

The function  $f(z)$  has simple poles at these points and the sequence of these poles converges toward the origin. The origin cannot be regarded as an isolated singularity because every one of its neighborhoods contains at least one pole (actually an infinite number of poles).

### 8.1.3 Weierstrass Theorem for Essential Singularities

The behavior of a function in the neighborhood of an isolated essential singularity is different from the cases of other isolated singularities such as poles and removable singularities. Most remarkable is the fact that  $f(z)$  can be made to take any arbitrary complex value by choosing an appropriate path of  $z \rightarrow a$ . For instance, if  $z$  approaches zero along the negative real semiaxis, then the function  $f(z) = e^{1/z}$  yields  $|f(z)| \rightarrow 0$ . However, if  $z$  approaches zero along the positive real semiaxis, then  $|f(z)| \rightarrow \infty$ . Finally, if  $z$  approaches zero along the imaginary axis, then  $|f(z)|$  remains constant but  $\arg f(z)$  oscillates, and so on. The character of a function near an essential singularity is described by the following theorem:

♠ **Weierstrass theorem:**

In any neighborhood of an isolated essential singularity, an analytic function approaches any given value arbitrarily closely.

*Proof* We use the contraposition method to prove our theorem. Let  $z = a$  be an isolated essential singularity of  $f(z)$ . We assume for the moment that for  $|z - a| < \varepsilon$ ,  $|f(z) - \gamma|$  with a given complex number  $\gamma$  does not become arbitrarily small. Then, the function  $[f(z) - \gamma]^{-1}$  is bounded in the region of  $|z - a| < \varepsilon$  so that it is possible to find a constant  $M$  such that

$$\left| \frac{1}{f(z) - \gamma} \right| < M \quad \text{for } |z - a| < \varepsilon.$$

Hence,  $[f(z) - \gamma]^{-1}$  is analytic for  $|z - a| < \varepsilon$  (or at worst has a removable singularity) and can be expanded by

$$\frac{1}{f(z) - \gamma} = b_0 + b_1(z - a) + b_2(z - a)^2 + \dots \quad (8.3)$$

If  $b_0 \neq 0$ , then

$$\lim_{z \rightarrow a} \frac{1}{f(z) - \gamma} = b_0 \text{ so that } \lim_{z \rightarrow a} f(z) = \gamma + \frac{1}{b_0}.$$

This means that  $z = a$  is not a singularity of  $f(z)$ , which contradicts our assumption. Otherwise, if  $b_0 = 0$ , we have

$$f(z) = \gamma + \frac{1}{(z - a)^k [b_k + b_{k+1}(z - a) + \cdots]},$$

where  $b_k$  is the first nonzero coefficient in the series (8.3). This clearly shows that  $z = a$  is a pole of  $f(z)$  of  $k$ th degree, which again is inconsistent with our assumption. Therefore, we conclude that  $|f(z) - \gamma|$  with a given  $\gamma$  can be arbitrarily small in the vicinity of an essential singularity  $z = a$ . Furthermore, since  $\gamma$  is arbitrary, the function  $f(z)$  approaches any given complex value arbitrarily closely. ♣

**Remark.** The above theorem becomes invalid if the **point at infinity** is taken into account; the point at infinity  $z = \infty$  is defined as the point  $\bar{z}$  that is mapped onto the origin  $z = 0$  by the transformation  $\bar{z} = 1/z$ . For instance, the function  $f(z) = e^z$  has an essential singularity at  $z = \infty$  but never approaches zero there.

#### 8.1.4 Rational Functions

In comparisons with the previous case, the behavior of an analytic function near a pole is easy to describe. We now derive the following result:

##### ♠ Theorem:

A **rational function** has no singularities other than poles. Conversely, an **analytic function** that has no singularities other than poles is necessarily a rational function.

A **rational function**  $f(z)$  is of the form

$$f(z) = \frac{p(z)}{q(z)}, \quad (8.4)$$

where

$$p(z) = \alpha_0 + \alpha_1 z + \alpha_2 z^2 + \cdots + \alpha_n z^n$$

and

$$q(z) = \beta_0 + \beta_1 z + \beta_2 z^2 + \cdots + \beta_m z^m.$$

Observe that the polynomials  $p(z)$  and  $q(z)$  are analytic at all finite points on the complex plane.

*Proof* In what follows, we assume that  $p(z)$  and  $q(z)$  have no common zeros; if they do have a common zero at  $z = z_0$ , it is always possible to write  $f(z)$  in (8.4) as the quotient of two polynomials with no common zeros by canceling a suitable number of the  $(z - z_0)$ -factors.

Obviously, the only possible singularities of  $f(z)$  are situated at the zeros of  $q(z)$ . Since the zeros of  $p(z)$  do not coincide with those of  $q(z)$ ,  $f(z)$  necessarily diverges at the zeros of  $q(z)$ . Such points can be poles but not essential singularities in view of the Weierstrass theorem given in Sect. 8.1.3. We have thus proved that all singularities of rational functions  $f(z)$  are necessarily poles.

To prove the converse, suppose that all the singularities of an analytic function  $f(z)$  are poles at the points  $a_1, a_2, \dots, a_n$ . The orders of these poles are denoted by  $m_1, m_2, \dots, m_n$ , respectively. In the vicinity of the point  $a_\nu$ , the function  $f(z)$  has a Laurent series expansion of the form

$$f(z) = \frac{c_{-m_\nu}^{(\nu)}}{(z - a_\nu)^{m_\nu}} + \dots + \frac{c_{-1}^{(\nu)}}{(z - a_\nu)} + \sum_{\mu=0}^{\infty} c_\mu^{(\nu)}(z - a_\nu)^\mu,$$

where the superscripts  $(\nu)$  on  $c^{(\nu)}$  indicate that they are the coefficients that belong to the  $\nu$ th poles,  $z = a_\nu$ . Denote the principal part by

$$g_\nu(z) = \frac{c_{-m_\nu}^{(\nu)}}{(z - a_\nu)^{m_\nu}} + \dots + \frac{c_{-1}^{(\nu)}}{(z - a_\nu)} \quad (8.5)$$

and consider the expression

$$h(z) = f(z) - g_1(z) - g_2(z) - \dots - g_n(z).$$

Since  $f(z) - g_\nu(z)$  is analytic at  $z = a_\nu$ , and  $g_\nu(z)$  is analytic everywhere except at  $z = a_\nu$ , it follows that  $h(z)$  is analytic at all points of the complex plane, including the **point at infinity**. In view of Liouville's theorem such a function is necessarily a constant. Thus we have identically  $h(z) \equiv \gamma_0$ , whence

$$f(z) = \gamma_0 + \sum_{\nu=1}^n g_\nu(z), \quad (8.6)$$

which implies that  $f(z)$  can be brought into the form (8.4). This completes the proof of our theorem. ♣

## Exercises

1. Find the poles and their order of the following functions:

$$(a) f(z) = \frac{\sin(z+1)}{z^3}, \quad (b) f(z) = \frac{\sin z}{z^3}.$$

**Solution:** (a) Clearly,  $\lim_{z \rightarrow 0} z^2 f(z) = \infty$  and  $\lim_{z \rightarrow 0} z^3 f(z) = \sin(1) \neq 0$ . Hence,  $f$  has a third-order pole at  $z = 0$  arising from the factor  $1/z^3$ . (b) Since  $\lim_{z \rightarrow 0} z^3 f(z) = 0$ , the pole of  $f(z)$  is not a third-order pole. Instead, noting the asymptotic behavior of  $\sin z$  near  $z = 0$ , we obtain

$$\lim_{z \rightarrow 0} z^2 f(z) = \lim_{z \rightarrow 0} z^2 \frac{z - (z^3/3!) + \cdots}{z^3} = 1.$$

Hence,  $f(z)$  has a second-order pole at  $z = 0$ . ♣

2. Show that a function  $f(z)$  cannot be bounded in the neighborhood of its isolated singular point  $z = a$ .

**Solution:** Use the contraposition method; if  $|f(z)| < M$  for  $|z - a| \leq r$ , then the expansion coefficients read

$$|c_{-n}| = \left| \frac{1}{2\pi i} \oint_C (\zeta - a)^{n-1} f(\zeta) d\zeta \right| \leq Mr^n \text{ for any } n,$$

where  $C$  is the circle given by  $|z - a| = r$ . Since  $r$  may be taken as small as desired, we have

$$c_{-1} = c_{-2} = \cdots = 0,$$

which means that the Laurent series reduces to a Taylor series.

Hence,  $f(z)$  should be analytic at  $z = a$ , which contradicts the assumption that  $z = a$  is a singular point. ♣

3. Let both  $f(z)$  and  $g(z)$  be analytic in the vicinity of  $z = a$  and have a zero of  $m$ th order at  $z = a$ . Prove that

$$\lim_{z \rightarrow a} \frac{f(z)}{g(z)} = \frac{f^{(m)}(a)}{g^{(m)}(a)}. \quad (8.7)$$

This result is called **l'Hôpital's rule**.

**Solution:** In the vicinity of  $z = a$ , we have

$$f(z) = (z - a)^m \left[ \frac{f^{(m)}(a)}{m!} + (z - a) \frac{f^{(m+1)}(a)}{(m+1)!} + (z - a)^2 \frac{f^{(m+2)}(a)}{(m+2)!} + \cdots \right],$$

and we also have a form similar to  $g(z)$ . These expressions immediately yield the desired equation (8.7). ♣

4. Prove that if  $f(z)$  has an essential singularity at  $z = a$ ,  $1/f(z)$  also has an essential singularity.

**Solution:** Suppose that  $f$  has an essential singularity at  $z = a$  but that  $1/f$  does not. If this is true,  $1/f$  will at most have a pole

there (of order  $N$ , for instance) and is expressed in terms of the series as

$$\frac{1}{f} = \sum_{n=-N}^{\infty} b_n h^n.$$

Rewrite this to obtain

$$f = \frac{h^N}{\sum_{m=0}^{\infty} b_{m-N} h^m}.$$

Note that the denominator  $\sum b_{m-N} h^m$  is analytic within  $C_1$ , and thus the fraction  $1/\sum b_{m-N} h^m$  is as well. As a result, the function  $f$  would be expanded into a power series in  $h$  starting with  $h^N$ ; this result contradicts our assumption that  $f(z)$  has an essential singularity at  $z = a$ . Therefore, wherever  $f(z)$  has an essential singularity,  $1/f$  also necessarily has one. ♣

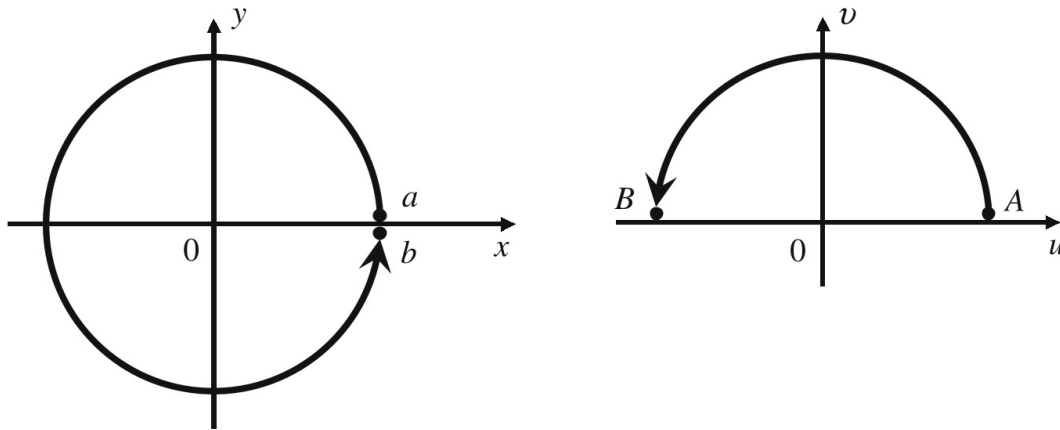
**Remark.** The above result sounds intriguing when compared with the behavior of an  $f(z)$  that has a pole. If  $f(z)$  has a pole of order  $N$  at  $z = a$ ,  $1/f$  obviously has no pole but does have a zero of order  $N$ ; i.e.,  $1/f \propto (z - a)^N$ .

## 8.2 Multivaluedness

### 8.2.1 Multivalued Functions

Up to this point, our concern has been limited to single-valued functions, i.e., functions whose values are uniquely specified once  $z$  is given. When we consider multivalued functions, many important theorems must be reformulated.

The necessary concepts are best illustrated by considering the behavior of the function  $f(z) = z^{1/2}$  in a graphical manner. Figure 8.1 gives a contour of a



**Fig. 8.1.** Mapping of a circle on the  $z$ -plane onto an upper-half circle on the  $w$ -plane through  $f(z) = z^{1/2}$

unit circle  $a \rightarrow b$  on the  $z$ -plane. Through the transformation  $w = f(z) = z^{1/2}$ , the circle is mapped onto a semicircle  $A \rightarrow B$  on the  $w$ -plane such that

$$\begin{aligned} z = 1 & \rightarrow w = 1, \\ z = i = e^{\pi i/2} & \rightarrow w = \left(e^{\pi i/2}\right)^{1/2} = e^{\pi i/4}, \\ z = -1 = e^{\pi i} & \rightarrow w = \left(e^{\pi i}\right)^{1/2} = e^{\pi i/2}, \\ z = -i = e^{3\pi i/2} & \rightarrow w = \left(e^{3\pi i/2}\right)^{1/2} = e^{3\pi i/4}. \end{aligned}$$

Of importance is the fact that the images of the points  $a$  and  $b$ , i.e.,  $A$  and  $B$ , respectively, are not equal but are distinct on the  $w$ -plane. This suggests that the value of  $z^{1/2}$  for  $z = 1$  is not uniquely determined. Furthermore, a similar phenomenon occurs for any circular contour  $a \rightarrow b$  with an arbitrarily large (or small) radius. We thus see that the function  $f(z) = z^{1/2}$  is multivalued, at least along the positive real axis; one point on the positive real axis of the  $z$ -plane is associated with two distinct points on the  $w$ -plane.

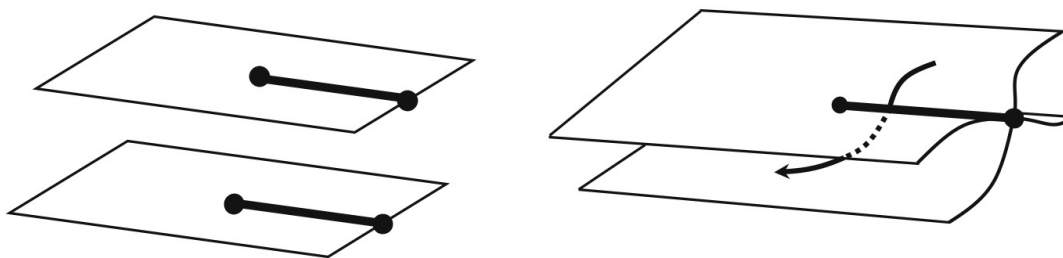
As a matter of fact, the multivaluedness of the function  $f(z) = z^{1/2}$  noted above occurs at all points on the whole  $z$ -plane (except at the origin). To see this, we observe again that the circular contour  $a \rightarrow b$  may have any radius. As a result, all the points on the  $z$ -plane are correlated with only half of the points on the  $w$ -plane, those for which  $\text{Im } [w] = v > 0$ . The remaining values of  $w$  are generated if a second circuit  $a \rightarrow b$  is made. Namely, the values of  $w$  with  $v < 0$  will be correlated with those values of  $z$  whose arguments lie between  $2\pi$  and  $4\pi$ . As a consequence, all values for  $z^{1/2}$  represented by on the  $w$ -plane may be divided into two independent sets: the set of values of  $w$  generated on the first circuit of the  $z$ -plane  $0 < \phi < 2\pi$  and those generated on the second circuit  $2\pi < \phi < 4\pi$ . These two independent sets of values for  $z^{1/2}$  are called the **branches** of  $z^{1/2}$ .

The concept of branch allows us to apply the theory of analytic functions to many-valued functions, where each branch is defined as a single-valued continuous function throughout its region of definition.

### 8.2.2 Riemann Surfaces

For the case  $z^{1/2}$ , the notion that the regions  $0 < \phi < 2\pi$  and  $2\pi < \phi < 4\pi$  correspond to two different regions of the  $w$ -plane is awkward geometrically, since each of these two regions covers the  $z$ -plane completely. To re-establish the single-valuedness and continuity of  $f(z)$ , it is desirable to give separate geometric meanings to two  $z$ -plane regions. This is achieved through the use of the notion of **Riemann surfaces**.

A Riemann surface is an ingenious device for representing both branches by means of a single continuous mapping. Suppose that two separate  $z$ -planes are cut along the positive real semiaxis from  $+\infty$  to  $0$  (see Fig. 8.2), and that the planes are superimposed on each other but retain their separate identities.



**Fig. 8.2.** A Riemann surface composed of two separated  $z$ -planes

Now suppose that the first quadrant of the upper sheet is joined along the cut to the fourth quadrant of the lower sheet to form a continuous surface. It is now possible to start a curve  $C$  in the first quadrant of the upper sheet, go around the origin, and cross the positive real semiaxis into the first quadrant of the lower sheet in a continuous motion. The curve can be continued on the lower sheet around the origin into the first quadrant of the lower sheet. This process of cutting and cross-joining two planes leads to the formation of a Riemann surface, which is thought of as a single continuous surface formed of two Riemann sheets.

Several important remarks are in order.

1. According to this model, the positive real semiaxis appears as a line where all four edges of our cuts meet. However, the Riemann surface has no such property. This results in the line between the first quadrant of the upper sheet and the fourth quadrant of the lower sheet being considered distinct from the line between the first quadrant of the lower sheet and the fourth quadrant of the upper one. There are two real positive semiaxes on the Riemann surface just as there are two real negative semiaxes. Hence, the entire Riemann surface is mapped one-to-one onto the  $w$ -plane. (The origin  $z = 0$  belongs to neither branch since the polar angle  $\theta$  is not defined for  $z = 0$ .)
2. The splitting of a multivalued function into branches is arbitrary to a great extent. For instance, we can define the following two functions, both of which may be treated as branches of  $f(z) = \sqrt{z}$ :

$$\text{Branch A : } f_A(z) = \begin{cases} \sqrt{r}e^{i\theta/2} & \text{for } 0 < \theta \leq \pi, \\ \sqrt{r}e^{i(\theta+2\pi)/2} & \text{for } -\pi < \theta \leq 0. \end{cases}$$

$$\text{Branch B : } f_B(z) = \begin{cases} \sqrt{r}e^{i(\theta+2\pi)/2} & \text{for } 0 < \theta \leq \pi, \\ \sqrt{r}e^{i\theta/2} & \text{for } -\pi < \theta \leq 0. \end{cases}$$

Note that branch A is continuous on the negative real semiaxis but is discontinuous on the positive real semiaxis (so is branch B). These two

branches together, constitute, the double-valued function  $f(z) = \sqrt{z}$ , and this representation is no better and no worse than the previous one.

3. The above-mentioned technique can be extended to other multivalued functions that require more than two Riemann sheets (for instance,  $f(z) = \sqrt[3]{z}$  requires three). There are functions requiring an infinite number of Riemann sheets, such as  $f(z) = z^\alpha$  with an irrational  $\alpha$ .

### 8.2.3 Branch Point and Branch Cut

We go back to the behavior of the multivalued function  $w = f(z) = z^{1/2}$  to introduce other important concepts referred to as **branch point** and **branch cut**. Let us consider a certain closed curve  $C$  without self-intersections in the  $z$ -plane. Specify a point  $z_0$  to which we assign a definite value of the argument  $\theta_0$ . Through the mapping  $w = z^{1/2}$ , we will find two distinct points:  $w_0(z_0)$  and  $w_1(z_0)$ .

In what follows, we examine the variation of the functions  $w_0(z)$  and  $w_1(z)$  as the point  $z$  moves continuously along the curve  $C$ . Since the argument of the point  $z$  on the curve  $C$  varies continuously, the functions  $w_0(z)$  and  $w_1(z)$  are continuous functions of  $z$  on the curve  $C$ .

Here, two different cases are possible. In the first case, the curve  $C$  does not contain the point  $z = 0$  within it. Then, after traveling the curve  $C$ , the argument of the point  $z_0$  returns to the original value  $\arg z_0 = \theta_0$ . Hence, the values of the functions  $w_0(z)$  and  $w_1(z)$  are also equal to their original values at the point  $z = z_0$  after traveling the curve  $C$ . Thus, in this case, two distinct single-valued functions of the complex variable  $z$  are defined on  $C$ :

$$w_0 = r^{1/2}e^{i\theta/2} \quad \text{and} \quad w_1 = r^{1/2}e^{i/2(\theta+2\pi)}.$$

Obviously, if the domain  $D$  of the  $z$ -plane has the property that any closed curve in the domain does not contain the point  $z = 0$ , then two distinct single-valued continuous functions,  $w_0(z)$  and  $w_1(z)$ , are defined in  $D$ . We call the functions  $w_0(z)$  and  $w_1(z)$  **branches** of the multivalued function  $w(z) = z^{1/2}$ .

In the second case, the curve  $C$  contains the point  $z = 0$  within it. Then, after traversing  $C$  in the positive direction, the value of the argument of the point  $z_0$  does not return to the original value  $\theta_0$  but changes by  $2\pi$  as expressed by

$$\arg z_0 = \theta_0 + 2\pi.$$

Therefore, as a result of their continuous variation after traversing the curve  $C$ , the values of the functions  $w_0(z)$  and  $w_1(z)$  at the point  $z_0$  are no longer be equal to the original values. More precisely, we obtain

$$\tilde{w}_0(z_0) = w_0(z_0)e^{i\pi} \quad \text{and} \quad \tilde{w}_1(z_0) = w_1(z_0)e^{i\pi},$$

which indicate that the function  $w_0(z)$  goes into the function  $w_1(z)$  and vice versa. This recurrence phenomenon stems from the fact that  $z = 0$  is the branch point of the multivalued function  $f(z) = z^{1/2}$ . A formal definition of branch point is given below.

♠ **Branch point:**

Suppose that several of branches of  $f(z)$  are analytic in the neighborhood of  $z = a$  but not at  $z = a$ . Then, the point  $z = a$  is a **branch point** if and only if  $f(z)$  passes from one of these branches to another when  $z$  moves along a closed circuit around  $z = a$ .

**Remark.** The point at infinity,  $z = \infty$ , is a branch of  $f(z)$  if and only if the origin is a branch point of  $f(1/z)$ .

It is important to note that the branch points for a given multivalued function, always occur pairwise so that they are connected by a simple curve called the **branch cut (cut or branch line)**. Branch cuts bound the regions within which the individual single-valued branches are defined. For instance, in the case of  $f(z) = z^{1/2}$ , the branch cut ran from the branch point at  $z = 0$  to another branch point at  $z = \infty$  along the positive real axis. It should be emphasized here that any curve joining the origin ( $z = 0$ ) and the point of infinity ( $z = \infty$ ) would have done just as well. For example, we could have used the negative real axis as the branch cut, for which the regions

$$-\pi < \phi < \pi \quad \text{and} \quad \pi < \phi < 3\pi$$

(instead of  $0 < \phi < 2\pi$  and  $2\pi < \phi < 4\pi$ ) serve as the defining regions for the first and second branch. On the  $w$ -plane, these two would correspond to  $\text{Re } v > 0$  and  $v < 0$ , respectively. We therefore may choose the branch cut that is most convenient for the problem at hand.

**Remark.** The choice of branches and branch cuts for a given multivalued function is not unique; however, the branch points and the number of branches are uniquely determined once a function is given.

## Exercises

1. Examine the multivaluedness of a logarithm function  $\ln z$ .

**Solution:** Expressing  $z$  in polar form,  $\ln z = \ln(re^{i\phi}) = \ln r + i\phi$ , and changing  $\phi$  by  $2\pi k$  results in

$$\ln z(r, \phi + 2\pi k) = \ln r + i(\phi + 2\pi k) = \ln z(r, \phi) + 2\pi i k. \quad (8.8)$$

It follows from (8.8) that there is no nonzero value of  $k$  for which  $\ln z(r, \phi + 2\pi k)$  and  $\ln z(r, \phi)$  are equal. Therefore, the logarithm function is an infinite-valued function. ♣

2. Evaluate  $\log e$ ,  $\log(-1)$ ,  $\log(1+i)$  according to the expression (8.8).

**Solution:**  $\log e = \log |e| + i \arg e = 1 + 2n\pi i$ ,  
 $\log(-1) = \log |-1| + i \arg(-1) = (2n+1)\pi i$ ,  
 $\log(1+i) = \log |1+i| + i \arg(1+i) = \frac{\log 2}{2} + (2n + \frac{1}{4})\pi i$ . ♣

3. Evaluate  $1^i$  and  $i^i$  according to the definition of power functions:  $z^a = e^{a \log z}$ , where  $z(\neq 0)$  and  $a$  are complex numbers.

**Solution:**  
 $1^i = e^{i \log 1} = e^{i \cdot 2n\pi i} = e^{-2n\pi}$ ,  
 $i^i = e^{i \log i} = e^{i(2n + \frac{1}{2})\pi i} = e^{-(2n + \frac{1}{2})\pi}$ . ♣

4. Show that a power function  $z^{m/n}$  with an irreducible rational number  $m/n$  ( $n \geq 2$ ) is an  $n$ -valued function.

**Solution:** The multiple values of  $z(r, \phi)^{m/n} = r^{m/n} e^{im\phi/n}$  are found by varying the integer  $k$  in the expression:

$$z(r, \phi + 2\pi k)^{m/n} = r^{m/n} e^{im\phi/n} e^{i2\pi km/n} = e^{i2\pi km/n} z(r, \phi)^{m/n}.$$

Substituting  $k = n$  yields

$$z(r, \phi + 2\pi n)^{m/n} = e^{i2\pi m} z(r, \phi)^{m/n} = z(r, \phi)^{m/n},$$

wherein  $e^{i2\pi m} = 1$  for arbitrary  $m \in \mathbf{N}$ . Hence, all multiple values of  $z^{m/n}$  at a given  $z$  are found with a value of  $k$  in the range  $0 \leq k \leq n-1$ . Since there are  $n$  different values of  $k$  in this range,  $z^{m/n}$  is an  $n$ -valued function. ♣

## 8.3 Analytic Continuation

### 8.3.1 Continuation by Taylor Series

It is often the case that a complex function is defined only in a limited region in the complex plane. For instance, a series representation of a function is of use only within its radius of convergence, but provides no direct information about the function outside this radius of convergence. An illustrative example is a function  $f(z)$  defined by

$$f(z) = 1 + z + z^2 + \cdots. \quad (8.9)$$

Obviously, this function is identified with  $1/(1-z)$  for  $|z| < 1$ , whereas it diverges for  $|z| > 1$  and thus is no longer equivalent to  $1/(1-z)$ . Nevertheless, a sophisticated technique makes it possible to identify the function  $f(z)$  given

in (8.9) with  $1/(1-z)$  even for the region  $|z| > 1$ . This technique, by which the defined region of a function is extended to an ‘uncultivated’ region, is called **analytic continuation**. The resultant function may often be defined by sequential continuation over the entire complex plane without reference to the original region of definition.

To see an actual process of analytic continuation, we suppose that a function  $f$  is given as a power series around  $z = 0$ , with a radius of convergence  $R$  and a singular point of  $f$  being on the **circle of convergence**. We show that it is possible to extend the function outside  $R$ . We first note that at any point  $z = a$  within the circle ( $|z| < R$ ), we can evaluate not only the value of the series but all its derivatives at that point as well because the function  $f$  is analytic and the series representation has the same radius of convergence. Therefore, we can obtain a Taylor series of  $f(z)$  around  $z = a$  as

$$f(z) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!} (z-a)^n. \quad (8.10)$$

The radius of convergence of this series is the distance to the nearest singular point, say  $z = z_s$  (see Fig. 8.3a). The resultant circle of convergence with radius  $R_0 = |z_s - z_0|$  is indicated by the solid circle in the figure. One may setup this process using a new point, e.g.,  $z = b$ , not necessarily within the original circle of convergence (see Fig. 8.3b), about which a new series such as (8.10) can be set up (see Fig. 8.3c). Continuing on in this way, it is apparently possible by means of such a series of overlapping circles to obtain values for  $f$  for every point in the complex plane excluding the singular points.

Our current discussion can be summarized as follows:

1. Let  $f(z)$  be defined by its Taylor series expansion around  $z = a$  within some circle  $|z - a| = r$ .
2. Specify a certain point  $z = b$  within the circle and evaluate  $f(b), f'(b), \dots$  to obtain a Taylor series of  $f(z)$  around  $z = b$ .
3. Observe that the latter series converges within a circle  $|z - b| = r'$  that intersects the first circle but may contain a region that is not within the first circle.
4. Specify again another point  $z = c$  within the circle  $|z - b| = r'$  and repeat the process described above.

### 8.3.2 Function Elements

We know that the term ‘analytic continuation’ refers to a method that allows us to extend the defining region of a complex function. Alternatively, this term can refer to the function that is newly found through analytic continuation of some other function. The formal definition is given below.

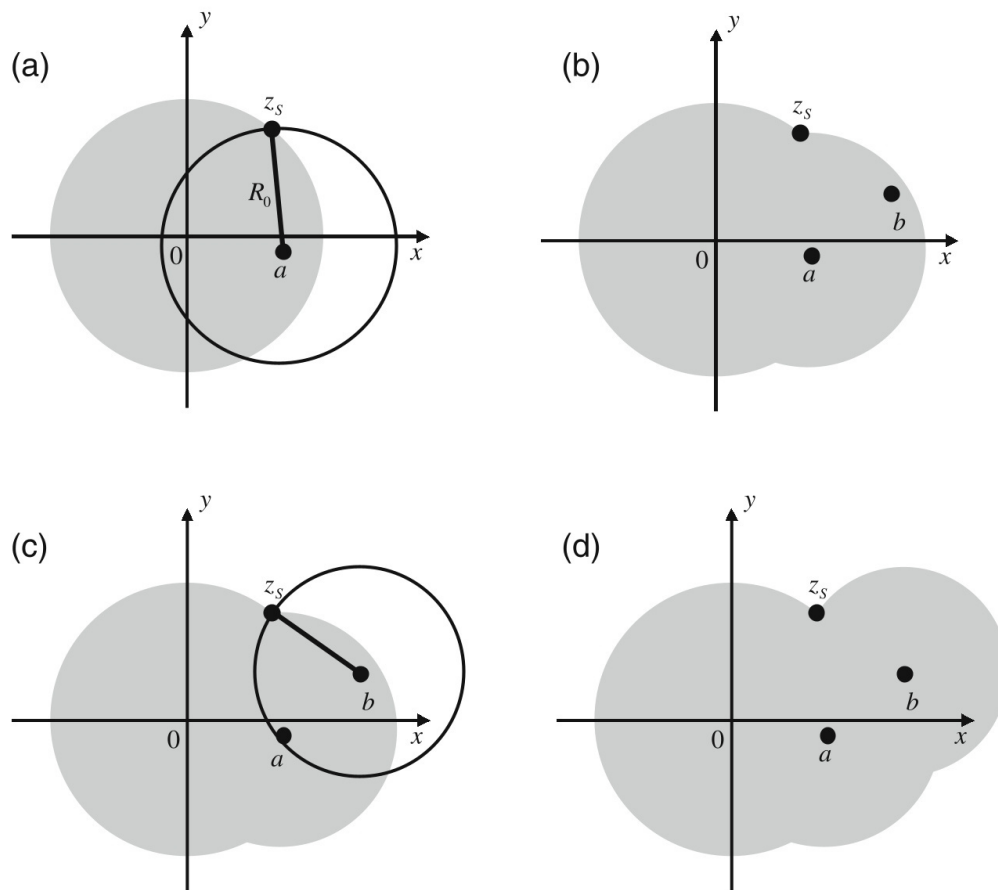


Fig. 8.3. Illustration of an analytic continuation procedure

♠ **Analytic continuation:**

Given a single-valued analytic function  $f_1(z)$  defined on a region  $D_1$ , the analytic function  $f_2(z)$  defined on  $D_2$  is called an **analytic continuation** of  $f_1(z)$  to  $D_2$  if and only if the intersection  $D_1 \cap D_2$  contains a simply connected open region where  $f_1(z) \equiv f_2(z)$ .

If the two analytic functions  $f_1(z)$  and  $f_2(z)$  defined on  $D_1$  and  $D_2$ , respectively, are analytic continuations of one another, then it is evident that an analytic function  $f(z)$  can be defined on  $D_1 \cup D_2$  by setting

$$f(z) = \begin{cases} f_1(z) & \text{in } D_1, \\ f_2(z) & \text{in } D_2. \end{cases}$$

Here,  $f_1$  and  $f_2$  are called **function elements** of  $f$ . More generally, we can consider a sequence of function elements  $(f_1, f_2, \dots, f_n)$  such that  $f_k$  is an analytic continuation of  $f_{k-1}$ . The elements of such a sequence are called **analytic continuations of each other**. Relevant terminology for this point is given below.

♠ **General analytic function:**

A general analytic function  $f$  is a nonvoid collection of function elements  $f_k$  in which any two elements are analytic continuations of each other by way of a chain whose links are members of  $f$ .

♠ **Complete analytic function:**

A complete analytic function  $f$  is a general analytic function that contains all the analytic continuations of any one of its elements.

A **complete analytic function** is evidently *maximal* in the sense that it cannot be further extended. Moreover, it is clear that every function element belongs to a unique complete analytic function. **Incomplete general analytic functions** are more arbitrary, and there are many cases in which two different collections of function elements should be regarded as defining the same function. For instance, a single-valued function  $f(z)$  defined in  $D$  can be identified either with the collection that consists of the single function element defined on  $D$  or with the collection of all function elements defined on  $D' \subset D$ .

*Examples* 1. Let us consider the functions

$$f_1(z) = \sum_{n=0}^{\infty} z^n \text{ defined on } |z| < 1 \quad (8.11)$$

and

$$f_2(z) = \sum_{n=0}^{\infty} \left(\frac{3}{5}\right)^{n+1} \left(z + \frac{2}{3}\right)^n \text{ defined on } \left|z + \frac{2}{3}\right| < \frac{5}{3}. \quad (8.12)$$

Both series converge to  $1/(1-z)$ ; Particularly the latter converges since

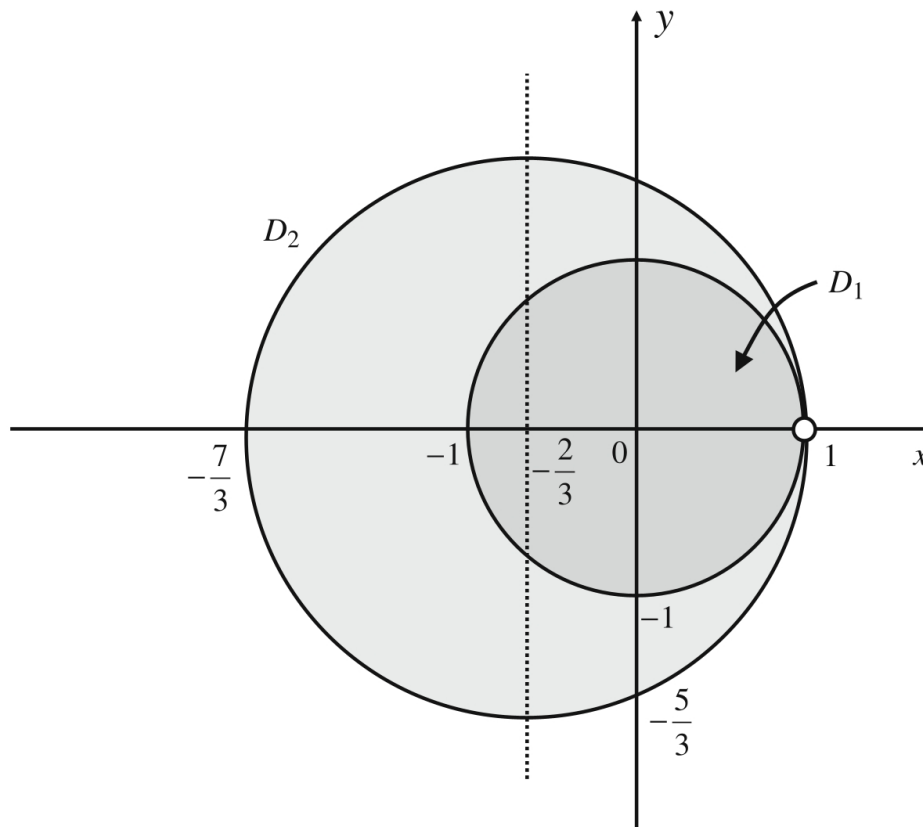
$$f_2(z) = \frac{3}{5} \sum_{n=0}^{\infty} \left[ \frac{3}{5} \left(z + \frac{2}{3}\right) \right]^n = \frac{\frac{3}{5}}{1 - \frac{3}{5}(z + \frac{2}{3})} = \frac{1}{1-z}.$$

Therefore, the two functions represent the same function  $f(z) = 1/(1-z)$  in the two overlapping regions (see Fig. 8.4), although they have different series representations. In this context, we can write

$$f(z) = \begin{cases} f_1(z) & \text{when for } z \in D_1, D_1 = \{z : |z| < 1\}, \\ f_2(z) & \text{when for } z \in D_2, D_2 = \{z : |z + \frac{2}{3}| < \frac{5}{3}\}. \end{cases}$$

2. Another illustrative example is given by

$$f_1(z) = \int_0^{\infty} e^{-zt} dt \text{ defined on } \operatorname{Re} z > 0 \quad (8.13)$$



**Fig. 8.4.** Both functions  $f_1(z)$  in (8.11) and  $f_2(a)$  in (8.12) represent the same function  $f(z) = 1/(1 - z)$  in the overlapping region  $D_1 \cap D_2$

and

$$f_2(z) = i \sum_{n=0}^{\infty} \left( \frac{z+i}{i} \right)^n \quad \text{defined on } |z+i| < 1.$$

Observe that each  $f_1$  and  $f_2$  reads  $1/z$  for the respective defining region. Thus, we have

$$\frac{1}{z} = \begin{cases} f_1(z) & \text{for } z \in D_1, \quad D_1 = \{z : \operatorname{Re} z > 0\}, \\ f_2(z) & \text{for } z \in D_2, \quad D_2 = \{z : |z+i| < 1\}. \end{cases}$$

The two functions are analytic continuations of one another, and  $f(z) = 1/z$  is the analytic continuation of both  $f_1$  and  $f_2$  for all  $z$  except  $z = 0$ .

**Remark.** In some cases, it is impossible to extend the function outside of a finite region because an infinite number of singularities are located densely on the boundary of the region. In that event, the boundary of this region is called the **natural boundary** of the function and the region within this boundary is called the **region of the existence** of the function.

### 8.3.3 Uniqueness Theorem

Having introduced the concept of analytic continuation, we may ask a question as to whether the function resulting from an analytic continuation process is uniquely determined, independent of the continuing path; i.e., whether a function that is continued along two different routes from one area to another will have the same value in the final area. We now attempt to answer this question by examining the theorem below.

♠ **Uniqueness theorem:**

Let  $f_1(z)$  and  $f_2(z)$  be analytic within a region  $D$ . If the two functions coincide in the neighborhood of a point  $z \in D$ , then they coincide throughout  $D$ .

*Proof* The theorem to be proven is rewritten in the following statement: *If both  $f(z)$  and  $g(z)$  are analytic at  $z_0$  and if  $f(z_n) = g(z_n)$  with  $n = 1, 2, \dots$  at points  $z_n$  that satisfy  $\lim_{n \rightarrow \infty} z_n = z_0$  but  $z_n \neq z_0$  for all  $n$ , then  $f(z) \equiv g(z)$  throughout  $D$ .* We now prove it.

Let  $h(z) = f(z) - g(z)$ . Here,  $f$  and  $g$  are assumed to satisfy the conditions given in the statement above, so that  $h(z_n) = 0$  for all  $n$  and  $h(z)$  is analytic at  $z_0$ . Owing to the analyticity of  $h(z)$  at  $z_0$ , we have the expansion

$$h(z) = a_0 + a_1(z - z_0) + a_2(z - z_0)^2 + \dots,$$

which converges in a certain circle around  $z_0$ . Since  $h(z)$  is continuous at  $z_0$ , we have

$$h(z_0) = \lim_{n \rightarrow \infty} h(z_n) = 0,$$

which means that the coefficient  $a_0$  is zero. Then, since  $h'(z)$  is also continuous at  $z_0$ , we set

$$h'(z_0) = \lim_{n \rightarrow \infty} h'(z_n) = 0,$$

which means that  $a_1 = 0$ . Continuing in this fashion, we find successively that all the coefficients vanish. In its circle of convergence, the function  $h(z)$  is therefore identically zero. This completes the proof. ♣

This remarkable theorem demonstrates the strong correlation between the behaviors of analytic functions on different parts of the complex plane. For example, if two functions agree in value over a small arc (arbitrarily small as long as it is not a point), then they are identical in their common region of analyticity.

### 8.3.4 Conservation of Functional Equations

An important consequence of the uniqueness theorem is the so-called principle of the conservation of a functional equation.

♠ **Conservation of functional equations:**

Let  $F(p, q, r)$  be an analytic function for all values of the three variables  $p, q, r$ , and let  $f(z)$  and  $g(z)$  be analytic functions of  $z$ . If a relation  $F[f(z), g(z), z] = 0$  between function elements  $f(z)$  and  $g(z)$  holds on a domain, then this relation is also true for all analytic continuations of these function elements.

**Remark.** In plain words, this theorem states that analytic continuations of  $f(z)$  satisfy every functional (and differential) equation satisfied by the original  $f(z)$ .

This theorem can easily be generalized to cases of functional equations involving more than two functions. We illustrate this by two examples.

*Examples 1.* From elementary trigonometry, we know that the real function  $\sin x$  has the additional theorem

$$\sin(x + u) = \sin x \cos u + \cos x \sin u,$$

where  $u$  is an arbitrary real value. Since  $\sin z$ ,  $\cos z$ , and  $\sin(z + u)$  are analytic for all finite values of  $z$ , and since the relation

$$\sin(z + u) = \sin z \cos u + \cos z \sin u$$

is satisfied if  $z$  is any point on the real axis, it follows by analytic continuation that the same relation must hold for all values of  $z$ . If we report the same argument with respect to the real variable  $u$ , we find that  $u$  may be replaced by a complex variable  $w$  without invalidating the relation in question. Hence, the **addition theorem** of the function  $\sin z$  is true for arbitrary complex values of  $z$  and  $w$ .

2. Another important example is afforded by functions satisfying differential equations. To take a simple case, we consider the function

$$f(z) = \log(1 + z).$$

This is represented for  $|z| < 1$  by the power series

$$f(z) = z - \frac{z^2}{2} + \frac{z^3}{3} - \cdots. \quad (8.14)$$

which yields

$$f'(z) = 1 - z + z^2 - z^3 + \cdots = \frac{1}{1 + z}.$$

In this context, the identity

$$f'(z) = \frac{1}{1+z} \quad (8.15)$$

appears to be valid for  $|z| < 1$ . However, it follows that the identity (8.15) must hold for all analytic continuations of the power series (8.14).

### 8.3.5 Continuation Around a Branch Point

The uniqueness theorem given in Sect. 8.3.3 also gives us the following corollary:

♠ **Theorem:**

If  $D_1$  and  $D_2$  are regions into which  $f(z)$  has been continued from  $D$ , yielding the corresponding functions  $f_1$  and  $f_2$ , and if  $D_3 = D_1 \cap D_2$  also overlaps  $D$ , then  $f_1 = f_2$  throughout  $D_3$ .

It is important to note that the validity of this theorem is due to the condition that  $D_3$  and  $D$  have a common region. If this condition is not satisfied, the uniqueness of analytic continuation may break down. Instead, one can say: If analytic continuation of a function  $f$  along two different routes from  $z_0$  to  $z_1$  yields two different values at  $z_1$ , then  $f(z)$  must have a certain kind of singularity between the two routes. This seems obvious by recalling the fact that the radius of convergence of a power series extends up to the next singularity of the function; if there were no singularities between the two routes, then it would be possible to fill in the region between the two routes by means of analytic continuation based on the power series. Then we would obtain sufficient overlapping so that the uniqueness theorem would be satisfied. In that event  $f(z_1)$  for the two different routes would be identical, in contradiction to our hypothesis. There must therefore be a singularity between the two routes.

Note that the last discussion does not state that different values must be obtained if there is any kind of singularity between the two routes. It must be a particular type of singularity to cause a discrepancy, and we call it a **branch point**, as we introduced earlier. An analytic function involving branch points is said to be multivalued and the various possible sets of values generated by the process of analytic continuation are known as **branches**. Intuitively, all the possible values of a function at a given point may be obtained by the process of analytic continuation if one winds about the branch point as many times as necessary.

### 8.3.6 Natural Boundaries

In all the examples considered so far, the singularities were isolated points. It is, however, easy to construct functions for which this is not the case. Consider, say, the function

$$f(z) = \frac{1}{\sin(1/z)}.$$

The denominator vanishes for  $1/z = n\pi$  with an integer  $n$ . Hence, the points  $z = (1/n\pi)$  are singular points of  $f(z)$ , but are clearly isolated in the vicinity of the origin. It is further possible for the singular points of a function to fill a whole arc of a continuous curve; in this case, we speak of a **singular line** of the function.

Particularly interesting is a situation in which a function  $f(z)$  has a closed singular line  $C$ . In this case, it is obviously impossible to continue  $f(z)$  analytically across  $C$ . The entire domain of definition of  $f(z)$  is therefore the interior of  $C$ , and we say that  $C$  is a **natural boundary** of  $f(z)$ .

Such an occurrence is not as unusual as it may seem. Consider, for instance, the analytic function  $f(z)$  defined by the power series

$$f(z) = z + z^2 + z^4 + z^8 + \cdots = \sum_{n=0}^{\infty} z^{2^n}. \quad (8.16)$$

By the root test given in Sect. 2.4.3, the circle of convergence of this series turns out to be  $|z| < 1$ . Thus  $f(z)$  must have at least one singularity on  $|z| = 1$ . For the sake of simplicity, we assume that this singularity is situated at the point  $z = 1$ ; a different location will cause a minor change in the argument. From the definition of  $f(z)$ , it follows that

$$f(z^2) = z^2 + z^4 + z^8 + \cdots = \sum_{n=1}^{\infty} z^{2^n} = f(z) - z.$$

By the principle of conservation (see Sect. 8.3.4), the functional equation

$$f(z) = z + f(z^2) \quad (8.17)$$

is true for all analytic continuations of  $f(z)$ . Observe that (8.17) gives

$$f'(z) = 1 + 2zf'(z^2),$$

which means that  $f(z)$  cannot have a derivative at  $z = -1$  since from hypothesis  $f(1)$  does not exist. Thus,  $z = -1$  is also a singular point of  $f(z)$ . In the same way, from the relation

$$f(z) = z + f(z^2) = z + z^2 + f(z^4)$$

it follows that the points  $z$  for which  $z^4 = 1$  are singularities of  $f(z)$ . Continuing in this fashion, we conclude that all points  $z$  for which  $z^{2^n} = 1$  are singularities of  $f(z)$ . But these are the points  $e^{2\pi i/(2^n)}$  that divide the circumference  $|z| = 1$  into  $2^n$  equal parts. Since, for  $n \rightarrow \infty$ , all points on  $|z| = 1$  are limits of these points and since the limit point of singular points is also a singularity, it follows that all points on  $|z| = 1$  are singular points of  $f(z)$ . We have thus proven that the unit circle is the natural boundary of the analytic function (8.16).

### 8.3.7 Technique of Analytic Continuations

The uniqueness theorem is the fundamental theorem in the theory of analytic continuation. However, in practice, the most relevant method would be one that tells us whether a function  $f_2$  is the analytic continuation of a function  $f_1$ .

Let us describe two possible methods of analytic continuation: The first is based on the **Schwarz principle of reflection**, which essentially makes use of the functional relation  $f(z^*) = f(z)^*$ .

♠ **Schwarz principle of reflection:**

If  $f(z)$  is analytic within a region  $D$  intersected by the real axis and is real on the real axis, then we have  $f(z^*) = f(z)^*$ .

*Proof* Expand  $f(z)$  in a Taylor series about a point  $a$  on the real axis. The coefficients of the Taylor series are real by virtue of the hypothesis that  $f(z)$  is real on the real axis. Hence, we have

$$f(z) = \sum_n c_n (z - a)^n, \quad (8.18)$$

where  $c_n$  is real. Then

$$f(z)^* = \sum_n c_n (z^* - a)^n = f(z^*), \quad (8.19)$$

proving the theorem. ♣

The above theorem holds for any point within the circle of convergence of the power series. By the methods of analytic continuation, therefore, it may be extended to include any nonsingular point conjugate to a point in  $D$ . As a result, the function in question can be continued from a region above the real axis to a region below.

A second method employs explicit functional relations such as **addition formulas** or **recurrence relations**. A simple example is provided by the addition formula

$$f(z + z_1) = f(z)f(z_1).$$

If  $f$  were known only in a given region, it would be continued outside that region to any point given by the addition of the coordinates of any two points within the region. A less trivial example occurs in the theory of gamma functions. The **gamma function** is defined by the integral

$$\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt. \quad (8.20)$$

This integral converges only for  $\operatorname{Re} z > 0$ , so that it defines  $\Gamma(z)$  for only the right half of the complex plane. From (8.20), one may readily derive (by integrating by parts) a functional relationship between  $\Gamma(z)$  and  $\Gamma(z+1)$ :

$$z\Gamma(z) = \Gamma(z+1). \quad (8.21)$$

We may now use (8.21) to continue  $\Gamma(z)$  into the  $\operatorname{Re} z < 0$  part of the complex plane. As first, we assume that  $\Gamma(z)$  is known for  $x > 0$ . Then using recurrence relation (8.21), the points in the strip  $-1/2 < x < 1/2$  can be computed in terms of the values of  $\Gamma(z)$  for  $x > 0$ . The function so defined and the original function have an overlapping region of convergence so that it is the analytic continuation into the negative  $x$ -region.

### 8.3.8 The Method of Moment

Suppose that we are given a power series  $f(z) = \sum_{n=0}^\infty a_n z^n$  where the coefficients  $a_n$  are the *moments* of a given continuous function. For example, suppose that there exists a continuous function  $g$  on  $[0, 1]$  such that

$$a_n = \int_0^1 g(t) t^n dt.$$

Then

$$f(z) = \sum_{n=0}^\infty \left[ \int_0^1 g(t) t^n dt \right] z^n = \sum_{n=0}^\infty \left[ \int_0^1 g(t) (zt)^n dt \right],$$

and interchanging the order of summation and integration, we find that

$$f(z) = \int_0^1 \left[ \sum_{n=0}^\infty g(t) (zt)^n \right] dt = \int_0^1 \frac{g(t)}{1-zt} dt.$$

(The interchange of summation and integration is easy to justify if  $|z| < 1$ .) Moreover, this integral form serves to define an analytic extension of the original power series.

*Examples* Consider

$$f(z) = \sum_{n=0}^\infty \frac{z^n}{n+1} \quad (|z| < 1). \quad (8.22)$$

Since

$$\frac{1}{n+1} = \int_0^1 t^n dt,$$

we set  $g(t) = 1$  to obtain

$$f(z) = \int_0^1 \frac{dt}{1-zt} \quad \text{for } |z| < 1.$$

The integral above is the analytic continuation of the original representation (8.22), so that the latter is analytic throughout the complex plane except for the semi-infinite line  $[1, \infty)$ . [In fact, the analytic continuation has a discontinuity at every point of the interval  $[1, \infty)$ .]

## Exercises

1. Suppose  $f(z) = \sum_{k=0}^{\infty} c_k z^{n_k}$  with  $\liminf_{k \rightarrow \infty} \frac{n_{k+1}}{n_k} > 1$ . Prove that the circle of convergence of  $f(z)$  above is a natural boundary for  $f$ .

**Solution:** Since the result is independent of  $c_k$ , we may assume without loss of generality that the radius of convergence is 1. In addition, neglecting finitely many terms if necessary, we assume that for some  $\delta > 0$  and for all  $k$ ,  $n_{k+1}/n_k = 1 + \delta$ . Finally, it suffices to show that  $f$  is singular at the point  $z = 1$ . The same result applied to the series  $\sum_{k=0}^{\infty} c_k (ze^{-i\theta})^{n_k}$  shows that  $f$  is singular at any point  $z = e^{i\theta}$ .

Choose an integer  $m > 0$  such that  $(m+1)/m < 1 + \delta$  and consider the power series  $g(w)$  obtained by setting  $z = (w^m + w^{m+1})/2$ . We then find that

$$\begin{aligned} g(w) &= f\left(\frac{w^m + w^{m+1}}{2}\right) \\ &= \frac{c_0}{2^{n_0}} w^{mn_0} + \frac{c_0 n_0}{2^{n_0}} w^{mn_0+1} + \cdots + \frac{c_0}{2^{n_0}} w^{mn_0+n_0} \\ &\quad + \frac{c_1}{2^{n_1}} w^{mn_1} + \frac{c_1 n_1}{2^{n_1}} w^{mn_1+1} + \cdots + \frac{c_1}{2^{n_1}} w^{mn_1+n_1} + \cdots \end{aligned}$$

Note that in this expression no two terms involve the same power of  $w$ , since the inequality  $mn_{k+1} > mn_k + n_k$  holds whenever  $n_{k+1}/n_k > (m+1)/m$ . If  $|w| < 1$ , then  $(|w|^m + |w|^{m+1})/2 < 1$ , and since  $f(z)$  is absolutely convergent for  $|z| < 1$ , the series  $\sum_{k=0}^{\infty} |c_k|[(|w|^m + |w|^{m+1})/2]^{n_k}$  converges. Hence, for  $|w| < 1$ ,  $g(w)$  is absolutely convergent. On the other hand, if we take  $w$  real and greater than 1, then  $(w^m + w^{m+1})/2 > 1$ , so the series  $\sum_{k=0}^{\infty} c_k[(w^m + w^{m+1})/2]^{n_k}$  diverges. Note, though, that the  $j$ th partial sums  $s_j$  of the above series are exactly the  $n_j(m+1)$ th partial sums of the power series of  $g$ . Hence, the series for  $g(w)$

diverges and  $g$ , too, has a radius of convergence of 1. This means that  $g(w)$  must have a singularity at some point  $w_0$  with  $|w_0| = 1$ . If  $w_0 \neq 1$ , then  $|(w^m + w^{m+1})/2| < 1$  and since  $f$  is analytic in  $|z| < 1$ ,  $g$  is analytic at  $w_0$ . Thus  $g$  must have a singularity at  $w_0 = 1$  and since  $g(w) = f[(w^m + w^{m+1})/2]$ ,  $f(z)$  must have a singularity at  $z = 1$ . ♣

2. Define an analytic continuation of: (i)  $\sum_{n=1}^{\infty} \frac{z^n}{\sqrt[3]{n}}$ , (ii)  $\sum_{n=0}^{\infty} \frac{z^n}{n^2 + 1}$ .

**Solution:**

(i) Since  $\frac{1}{n^{1/3}} = \Gamma(1/3) \int_0^{\infty} e^{-nt} t^{-2/3} dt$ , we have

$$\begin{aligned} \sum_{n=1}^{\infty} \frac{z^n}{n^{1/2}} &= \Gamma\left(\frac{1}{3}\right) \int_0^{\infty} \sum_{n=1}^{\infty} (ze^{-t})^n t^{-2/3} dt \\ &= \Gamma\left(\frac{1}{3}\right) \int_0^{\infty} \frac{z}{t^{2/3}(e^t - z)} dt, \end{aligned}$$

which is analytic outside of the interval  $[1, \infty)$ .

(ii) Since  $\frac{1}{n^2 + 1} = \int_0^{\infty} e^{-nt} \sin t dt$ ,

$$\sum_{n=0}^{\infty} \frac{z^n}{n^2 + 1} = \int_0^{\infty} \sum_{n=0}^{\infty} (ze^{-t})^n \sin t dt = \int_0^{\infty} \frac{e^t \sin t}{e^t - z} dt,$$

which is analytic outside of the interval  $[1, \infty)$ . ♣

3. Suppose that  $f$  is bounded and analytic in  $\text{Im} z \geq 0$  and real on the real axis. Prove that  $f$  is constant.

**Solution:** By the Schwarz reflection principle,  $f$  can be extended to the entire plane and would then be a bounded entire function.

Hence,  $f$  is constant. ♣

4. Given an entire function that is real on the real axis and imaginary on the imaginary axis, prove that it is an odd function; i.e.,  $f(z) = -f(-z)$ .

**Solution:** Set  $f(z) = f(x, y) = u(x + iy) + iv(x + iy)$ . The Schwarz reflection principle implies that  $f(z^*) = f(x - iy) = u(x - iy) + iv(x - iy) = u(x + iy) - iv(x + iy) = -f(z)$ . In a similar way, we have  $f(-z) = f(-x - iy) = u(-x - iy) + iv(-x - iy) = -u(x + iy) - iv(x + iy) = -f(z)$ . ♣



## Contour Integrals

**Abstract** In this chapter, we show that singularities do not interfere with the analysis of complex functions but are useful in extracting complex integrals along closed contours. This utility of singularities is based on the residue theorem (Sect. 9.1.1), argument principle (Sect. 9.4), and principal value integrals (Sect. 9.5.1), all of which correlate the nature of singularities within and/or on the contour with the relevant complex integrals.

### 9.1 Calculus of Residues

#### 9.1.1 Residue Theorem

In the preceding two chapters, we provided the theoretical bases of complex functions. This chapter deals with more practical matters that are relevant to computations of contour integrations on a complex plane. The theorem below is central to the development of this topic.

♠ **Residue theorem:**

If a function  $f(z)$  is analytic everywhere within a closed contour  $C$  except at a finite number of poles, its contour integral along  $C$  yields

$$\oint_C f(z)dz = 2\pi i \sum_j \text{Res}(f, a_j). \quad (9.1)$$

Here,  $\text{Res}(f, a_j)$  is called the **residue** of  $f(z)$  at the pole  $z = a_j$ . When the pole is  $m$ th order, it reads

$$\text{Res}(f, a_j) = \frac{1}{(m-1)!} \lim_{z \rightarrow a_j} \frac{d^{(m-1)}}{dz^{(m-1)}} [(z - a_j)^m f(z)]. \quad (9.2)$$

Once the residue is evaluated, the integral  $\oint_C f(z)dz$  around the contour  $C$  surrounding the pole  $z = a$  can be determined by the above theorem. Notably, this theorem enables us to evaluate various kinds of integrals of real functions that are unfeasible by means of elementary calculus.

Before demonstrating the utility of the residue theorem, we present a short review of the nature of residues. Originally, the residue of  $f(z)$  is defined in association with a particular coefficient of the Laurent series expansion. We know that  $f(z)$  around its pole at  $z = a$  may be expressed by a Laurent series expansion such as

$$f(a+h) = \sum_{n=-\infty}^{\infty} c_n h^n, \quad c_n = \frac{1}{2\pi i} \oint_C \frac{f(z)}{(z-a)^{n+1}} dz.$$

Then, the specific coefficient

$$c_{-1} = \frac{1}{2\pi i} \oint_C f(z) dz \quad (9.3)$$

is called the residue of  $f(z)$  at  $z = a$ . In fact, the result (9.3) immediately reduces to the form of (9.1) as

$$\frac{1}{2\pi i} \oint_C f(z) dz = 2\pi i c_{-1}.$$

The equivalence of the two quantities,  $\text{Res}(f, a)$  in (9.2) and  $c_{-1}$  in (9.3), is verified as follows.

*Proof (of the residue theorem).* Suppose that  $f(z)$  has a pole of order  $m$  at  $a$ . Then  $f(z)$  can be written as

$$f(z) = \frac{c_{-m}}{(z-a)^m} + \frac{c_{-m+1}}{(z-a)^{m-1}} + \cdots + \frac{c_{-1}}{(z-a)} + \sum_{n=0}^{\infty} c_n (z-a)^n. \quad (9.4)$$

Now we introduce the quantity

$$\begin{aligned} g(z) &\equiv (z-a)^m f(z) = c_{-m} + c_{-m+1}(z-a) + \cdots \\ &= \sum_{n=0}^{\infty} c_{n-m} (z-a)^n. \end{aligned} \quad (9.5)$$

Since  $g(z)$  is analytic everywhere in a neighborhood around  $a$ , it can be expanded in terms of a Taylor series as

$$g(z) = \sum_{n=0}^{\infty} \frac{g^{(n)}(a)}{n!} (z-a)^n. \quad (9.6)$$

The residue  $c_{-1}$  is the coefficient of the  $n = m-1$  term in (9.5). Hence, comparing (9.5) with (9.6), we have

$$c_{-1} = \frac{1}{(m-1)!} g^{(m-1)}(a) = \frac{1}{(m-1)!} \lim_{z \rightarrow a} \frac{d^{(m-1)}}{dz^{(m-1)}} [(z-a)^m f(z)], \quad (9.7)$$

which is simply equation (9.2). ♣

### 9.1.2 Remarks on Residues

The reason that only the particular coefficient  $c_{-1}$  plays a role in evaluating the contour integral is clarified by integrating both sides of (9.4) along the contour containing the  $m$ th-order pole  $a$ . For convenience, we rewrite (9.4) as

$$f(z) = \sum_{n=1}^m \frac{c_{-n}(a)}{(z-a)^n} + \Psi_A(z), \quad (9.8)$$

where

$$\Psi_A(z) = \sum_{n=0}^{\infty} c_n(a)(z-a)^n$$

is the regular part of the series (9.8), thus being analytic everywhere in a region within a closed contour  $C$  containing  $a$ . By integrating  $f(z)$  along the contour  $C$ , we set

$$\oint_C f(z) dz = \sum_{n=1}^m c_{-n} \oint_C \frac{1}{(z-a)^n} dz \quad (9.9)$$

because of the analyticity of  $\Psi_A(z)$ . The integral of (9.9) can be easily evaluated by letting the contour be a circle of radius  $\rho$  centered at  $a$ . Since any point on the contour can be expressed as  $z = a + \rho e^{i\phi}$ , we have

$$\oint_C \frac{1}{(z-a)^n} dz = \int_0^{2\pi} \frac{i\rho e^{i\phi}}{\rho^n e^{in\phi}} d\phi = i\rho^{-(n-1)} \int_0^{2\pi} e^{-i(n-1)\phi} d\phi. \quad (9.10)$$

Note that the integral (9.10) vanishes for all  $n \neq 1$ , and it is only when  $n = 1$  that it has a nonzero value:

$$\oint_C \frac{1}{z-a} dz = i \int_{\phi_0}^{\phi_0+2\pi} d\phi = 2\pi i.$$

Therefore, all the terms in the sum of (9.9) are zero except the  $n = 1$  term, and Goursat's formula takes the form

$$\oint_C f(z) dz = 2\pi i c_{-1}. \quad (9.11)$$

In short, once we integrate the function  $f(z)$  in (9.8), only the term involving  $c_{-1}$  survives, whereas the other terms vanish. This results in the fact that the contour integral  $\oint_C f(z) dz$  around a pole is determined by the value of the specific coefficient  $c_{-1}$ .

### 9.1.3 Winding Number

To evaluate  $\oint_C f(z)dz$  when  $C$  is a general closed curve (and when  $f$  may have isolated singularities), we introduce the following concept:

♠ **Winding number:**

Suppose that  $C$  is a closed curve and that the point  $z = a$  is not located on  $C$ . Then the number

$$n(C, a) = \frac{1}{2\pi i} \oint_C \frac{dz}{z - a}$$

is called the **winding number** of  $C$  around  $a$ .

Note that if  $C$  represents the boundary of a circle (traversed counterclockwise), then the winding number reads

$$n(C, a) = \begin{cases} 0 & \text{if } a \text{ is inside the circle,} \\ 1 & \text{if } a \text{ is outside the circle.} \end{cases}$$

Both identities have already been proven in the context of Cauchy's theorem. In addition, if the curve  $C$  encloses  $k$  times the point  $a$ , then we have

$$n(C, a) = \frac{1}{2\pi i} \int_0^{2k\pi} i d\theta = k,$$

which explains the terminology “winding number.”

♠ **Theorem:**

For any closed curve  $C$  and point  $a \notin C$ , the winding number  $n(C, a)$  is an integer.

*Proof* Suppose that  $C$  is parametrized by  $z(t)$ ,  $0 \leq t \leq 1$ , and set

$$f(s) = \int_0^s \frac{z'(t)}{z(t) - a} dt \quad (0 \leq s \leq 1).$$

Then, it follows from

$$f'(s) = \frac{z'(s)}{z(s) - a}$$

that the quantity

$$[z(s) - a]e^{-f(s)}$$

is a constant, and setting  $s = 0$ , we have

$$[z(s) - a]e^{-f(s)} = z(0) - a.$$

Hence,

$$e^{f(s)} = \frac{z(s) - a}{z(0) - a} \quad \text{and} \quad e^{f(1)} = \frac{z(1) - a}{z(0) - a} = 1,$$

since  $C$  is closed, i.e.,  $z(1) = z(0)$ . Thus

$$f(1) = 2\pi ki \quad \text{for some integer } k$$

and

$$n(C, a) = \frac{1}{2\pi i} f(1) = k. \quad \clubsuit$$

In terms of the winding number, the residue theorem given in Sect. 9.1.1 can be restated as follows:

♠ **Residue theorem (restated):**

Suppose  $f(z)$  is analytic in a simply connected domain  $D$  except for isolated singularities at  $z_1, z_2, \dots, z_m$ . Let  $C$  be a closed curve that does not intersect any of the singularities. Then

$$\oint_C f(z) dz = 2\pi i \sum_{k=1}^m n(C, z_k) \text{Res}(f, z_k). \quad (9.12)$$

The proof is left to the reader.

### 9.1.4 Ratio Method

We saw in Sect. 7.4.5 that a function having a pole of order  $m$  can be expressed by the ratio of two polynomials such as

$$f(z) = \frac{p(z)}{q(z)}. \quad (9.13)$$

In this case, it is possible to formulate an alternative equation that determines the residue of  $f(z)$ . Employing such an equation to evaluate the residue is referred to as a **ratio method**.

To derive these equations, we first recall the fact that if a function  $R(z)$  satisfies

$$p(a) = p'(a) = \dots = p^{(m-1)}(a) = 0 \quad \text{and} \quad p^{(m)}(a) \neq 0,$$

the Taylor series for  $R(z)$  is given by

$$p(z) = \frac{p^{(m)}(a)}{m!} (z - a)^m + h.o.,$$

where *h.o.* means the terms of higher order. Such a function, for which the lowest power of  $(z - a)$  is  $m$ , is said to have an ***m*th-order zero** at  $a$ .

Now we present the equation for the residue of  $f(z)$  at a simple pole  $a$ . As seen from (9.13), a simple pole of  $f(z)$  arises from the fact that  $p(z)$  has a zero of  $(m - 1)$ th order and  $q(z)$  has a zero of order  $m$ . Then,

$$f(z) = \frac{\frac{p^{(m-1)}(a)}{(m-1)!}(z-a)^{m-1} + h.o.}{\frac{q^{(m)}(a)}{m!}(z-a)^m + h.o.}.$$

For such a function, we obtain the residue of  $f$  at the simple pole  $a$  as

$$c_{-1} = \lim_{z \rightarrow a} (z - a)f(z) = m \frac{p^{(m-1)}(a)}{q^{(m)}(a)}. \quad (9.14)$$

By means of 9.14, we can compute the residue of  $f(z)$  at a simple pole  $a$  quite easily.

Next we consider the equation for a second-order pole of  $f(z)$  at  $a$ . Such a pole arises when  $p(z)$  has a zero of order  $m$  and  $q(z)$  has a zero of order  $(m + 2)$  at  $a$ . Then,

$$f(z) = \frac{\frac{p^{(m)}(a)}{m!}(z-a)^m + \frac{p^{(m+1)}(a)}{(m+1)!}(z-a)^{m+1} + h.o.}{\frac{q^{(m+2)}(a)}{(m+2)!}(z-a)^{m+2} + \frac{q^{(m+3)}(a)}{(m+3)!}(z-a)^{m+3} + h.o.},$$

from which we set

$$\begin{aligned} c_{-1} &= \lim_{z \rightarrow a} \frac{d}{dz} [(z - a)^2 f(z)] \\ &= \frac{m+2}{m+3} \cdot \frac{(m+3)p^{(m+1)}(a)q^{(m+2)}(a) - (m+1)p^{(m)}(a)q^{(m+3)}(a)}{[q^{(m+2)}(a)]^2}. \end{aligned} \quad (9.15)$$

For example, if the second-order pole of a function arises from a second-order zero of  $q(z)$ , then  $m = 0$ . The residue of such a pole is given by (9.15) as

$$c_{-1} = \frac{2}{3} \frac{3p'(a)q''(a) - p(a)q^{(3)}(a)}{[q''(a)]^2}. \quad (9.16)$$

### 9.1.5 Evaluating the Residues

In what follows, we demonstrate actual procedures to evaluate the residue by means of the three methods discussed in the previous subsections. As an instructive example, we consider the function

$$f(z) = \frac{e^z}{z(z+2)^2},$$

which has a simple pole at  $z = 0$  and a second-order pole at  $z = -2$ .

### Using a Laurent expansion:

The present purpose is to evaluate the coefficient  $c_{-1}$  of the Laurent series expansion of  $f(z)$  around the poles at  $z = 0$  and  $z = -2$ . In order to do this we first determine the Taylor series for the factor  $e^z/(z+2)^2$  around  $z = 0$ . Since the expressions

$$e^z = 1 + z + \frac{z^2}{2!} + \cdots$$

and

$$\frac{1}{(z+2)^2} = \frac{1}{4} \left[ \frac{1}{1+(z/2)} \right]^2 = \frac{1}{4} \left( 1 - z + \frac{3}{4}z^2 - \cdots \right)$$

hold around  $z = 0$ , we have

$$\frac{e^z}{z(z+2)^2} = \frac{1}{4z} \left( 1 + z + \frac{z^2}{2!} + \cdots \right) \left( 1 - z + \frac{3}{4}z^2 - \cdots \right) = \frac{1}{4z} + \frac{z}{16} + \cdots.$$

Thus, we immediately obtain

$$c_{-1}(0) = \frac{1}{4}.$$

Similarly we have

$$c_{-1}(-2) = -\frac{3}{4}e^{-2} \quad (\text{see Exercise 1}).$$

### Using Goursat's formula:

The residue of the simple pole at  $z = 0$  is given by

$$c_{-1}(0) = \lim_{z \rightarrow 0} \left[ z \frac{e^z}{z(z+2)^2} \right] = \frac{1}{4}$$

and that of the second-order pole at  $z = -2$  is given by

$$c_{-1}(-2) = \frac{1}{1!} \lim_{z \rightarrow -2} \frac{d}{dz} \left[ (z+2)^2 \frac{e^z}{z(z+2)^2} \right] = -\frac{3}{4}e^{-2}.$$

**Using the ratio method:**

For this example, the numerator and denominator functions can be chosen in different ways. For the residue at  $z = 0$ , we could take

$$p(z) = e^z, \quad q(z) = z(z+2)^2$$

or, alternatively,

$$p(z) = \frac{e^z}{(z+2)^2}, \quad q(z) = z.$$

For either choice, the residue for the simple pole is given by

$$c_{-1}(0) = \frac{p(0)}{q'(0)} = \frac{1}{4}.$$

The residue  $c_{-1}(-2)$  can be obtained in a similar manner as above (see Exercise 2).

**Exercises**

1. Evaluate the residue of

$$f(z) = \frac{e^z}{z(z+2)^2}$$

at  $z = -2$  by using a Laurent expansion.

**Solution:** The residue of  $f(z)$  at  $z = -2$  is found by using the expression

$$e^z = e^{-2}e^{z+2} = e^{-2} \sum_{n=0}^{\infty} \frac{(z+2)^n}{n!} = e^{-2} \left[ 1 + (z+2) + \frac{(z+2)^2}{2!} + \cdots \right],$$

and the Taylor series expansion for  $1/z$  around  $z = -2$  as

$$\frac{1}{z} = -\frac{1}{2} \left[ \frac{1}{1 - (z+2)/2} \right] = -\sum_{m=0}^{\infty} \frac{(z+2)^m}{2^{m+1}} = -\frac{1}{2} - \frac{z+2}{4} - \frac{(z+2)^2}{8} - \cdots.$$

Thus, the Laurent series for  $f(z)$  around  $z = -2$  is

$$\frac{e^z}{z(z+2)^2} = -\frac{1}{2}e^{-2} \left[ \frac{1}{(z+2)^2} + \frac{3}{2} \frac{1}{z+2} + \frac{5}{4} + \cdots \right],$$

from which we have

$$c_{-1}(-2) = -\frac{3}{4}e^{-2}. \quad \clubsuit$$

2. Evaluate the residue of

$$f(z) = \frac{e^z}{z(z+2)^2}$$

at  $z = -2$  using the ratio method.

**Solution:** For the pole at  $z = -2$ , we can choose either

$$p(z) = e^z, \quad q(z) = z(z+2)^2$$

as before or

$$p(z) = \frac{e^z}{z}, \quad q(z) = (z+2)^2.$$

Then, regardless of how the numerator and denominator are chosen, we refer to (9.16) to obtain

$$c_{-1}(-2) = \frac{2}{3} \frac{3p'(-2)q''(-2) - p(-2)q^{(3)}(-2)}{[q''(-2)]^2} = -\frac{3}{4}e^{-2}. \quad \clubsuit$$

## 9.2 Applications to Real Integrals

### 9.2.1 Classification of Evaluable Real Integrals

Using the residue theorem, we can evaluate the five types of real integrals listed below.

1.  $\int_0^{2\pi} f(\cos \theta, \sin \theta) d\theta$ , where  $f(x, y)$  is a rational function without a pole on the circle  $x^2 + y^2 = 1$ .
2.  $\int_{-\infty}^{\infty} f(x) dx$ , where  $f(z)$  is a rational function without a real pole and is subject to the condition that  $\lim_{|x| \rightarrow \infty} xf(x) = 0$ .
3.  $\int_{-\infty}^{\infty} f(x)e^{ix} dx$ , where  $f(z)$  is an analytic function in the upper-half plane  $\text{Im} z \geq 0$  except at a finite number of points.
4.  $\int_0^{\infty} f(x)/x^\alpha dx$ , where  $\alpha$  denotes a real number such that  $0 < \alpha < 1$  and  $f(z)$  is a rational function with no pole on the positive real axis  $x \geq 0$ , which satisfies the condition  $f(z)/z^{\alpha-1} \rightarrow 0$  as  $z \rightarrow 0$  and  $z \rightarrow \infty$ .
5.  $\int_0^{\infty} f(x) \log x dx$ , where  $f(z)$  is a rational function with no pole on the positive real axis  $x \geq 0$  and satisfies the condition  $\lim_{x \rightarrow +\infty} xf(x) = 0$ .

In Sect. 9.2.2–9.2.6 we demonstrate actual processes for evaluating the above integrals.

### 9.2.2 Type 1: Integrals of $f(\cos \theta, \sin \theta)$

Consider an integral of the form

$$\int_0^{2\pi} f(\cos \theta, \sin \theta) d\theta.$$

Setting  $z = e^{i\theta}$  makes it a contour integral around the unit circle, and thus the evaluation of the residues within the circle completes the integration.

*Example* We evaluate the integral

$$I = \int_0^{2\pi} \frac{d\theta}{1 - 2p \cos \theta + p^2} \quad (p < 1). \quad (9.17)$$

If we express  $\cos \theta$  in terms of  $z = e^{i\theta}$ , 9.17 becomes a contour integral,

$$I = \oint_C \frac{1}{1 - p(z + z^{-1}) + p^2} \frac{dz}{iz} = \frac{1}{i} \oint_C \frac{dz}{(1 - pz)(z - p)}, \quad (9.18)$$

where  $C$  is a unit circle centered at the origin. The integrand in (9.18) has a simple (first-order) pole at  $z = p$  within  $C$ . Hence, we obtain

$$I = \frac{1}{i} \times 2\pi i \lim_{z \rightarrow p} \left( \frac{1}{1 - pz} \right) = \frac{2\pi}{1 - p^2}. \quad \clubsuit$$

### 9.2.3 Type 2: Integrals of Rational Function

Next consider the integral

$$I = \int_{-\infty}^{\infty} f(x) dx, \quad (9.19)$$

where  $f(x)$  is a rational function subject to the condition

$$\lim_{|x| \rightarrow \infty} x f(x) = 0,$$

which is a necessary and sufficient condition for the integral to be convergent. To evaluate (9.19), we consider the integral of  $f(z)$  along a closed contour consisting of the real axis from  $-R$  to  $+R$  and a semicircle  $\Gamma(R)$  in the upper half-plane. The contour integral is expressed as

$$\oint_C f(z) dz = \int_{-R}^R f(x) dx + \int_{\Gamma(R)} f(z) dz. \quad (9.20)$$

From the Lemma below, it follows that the second term in (9.20) vanishes in the limit  $R \rightarrow \infty$ . Hence, we obtain

$$\lim_{R \rightarrow \infty} \oint_C f(z) dz = \int_{-\infty}^{\infty} f(x) dx, \quad (9.21)$$

and applying the residue theorem yields

$$\int_{-\infty}^{\infty} f(x) dx = 2\pi i \sum_j \text{Res}(f, a_j),$$

where  $a_j$  is the  $j$ th pole of  $f(z)$  in the upper half-plane. Therefore, the evaluation of the residues located within the upper half-plane completes the integration.

*Example* We prove the equation

$$I = \int_{-\infty}^{\infty} \frac{dx}{1+x^2} = \pi.$$

Since  $x/(1+x^2)$  vanishes as  $|x| \rightarrow \infty$ , we may follow a process similar to the one discussed above. Since  $z = i$  is the only pole of  $1/(1+z^2) = 1/(z+i)(z-i)$  involved in the upper half-plane, we have

$$I = (2\pi i) \cdot \text{Res}(i) = 2\pi i \frac{1}{2i} = \pi.$$

Less simple examples will be found in Exercises Sect. 9.2. ♣

As was noted earlier, our result (9.21) is based on the following lemma:

♠ **Lemma:**

Let  $f(z)$  be continuous in the sector  $\theta_1 < \arg z < \theta_2$ . If

$$\lim_{|z| \rightarrow \infty} z f(z) = 0 \text{ for } \theta_1 < \arg z < \theta_2, \quad (9.22)$$

then the integral  $\int f(z) dz$  extended over the arc of the circle  $|z| = r$  contained in the sector tends to 0 as  $r \rightarrow \infty$ .

*Proof* Let  $M(r)$  be the upper bound of  $|f(z)|$  on the arc of the circle  $|z| = r$ . Then we have

$$\left| \int f(z) dz \right| \leq M(r) \int_{\theta_1}^{\theta_2} r d\theta = M(r) \cdot r(\theta_2 - \theta_1). \quad (9.23)$$

In view of the condition (9.22), the right-hand side of (9.23) vanishes as  $r \rightarrow \infty$ . This completes the proof. ♣

### 9.2.4 Type 3: Integrals of $f(x)e^{ix}$

We now study integrals of the form

$$\int_{-\infty}^{\infty} f(x)e^{ix} dx,$$

where  $f$  is analytic on the upper half-plane  $\text{Im} z \geq 0$  except at a finite number of singularities (if they exist). We first consider the case when the singularities are not on the real axis. Then, the integral

$$\int_{-R}^R f(x)e^{ix} dx$$

has a meaning, which can be seen from the following theorem:

♠ **Theorem:**

If  $\lim_{|z| \rightarrow \infty} f(z) = 0$  for  $\text{Im} z \geq 0$ , then

$$\lim_{R \rightarrow +\infty} \int_{-R}^R f(x)e^{ix} dx = 2\pi i \sum \text{Res} [f(z)e^{ix}],$$

the summation extending over the singularities of  $f(z)$  contained in the upper half-plane  $y > 0$ .

Before starting the proof, we note that  $|e^{iz}| \leq 1$  in the half-plane  $y \geq 0$ . This leads us to integrate on the half-plane  $y \geq 0$  along the contour used above for an integral of type **2**. To prove the theorem, thus it suffices to show that the integral  $\int_{\Gamma(R)} f(z)e^{iz} dz$  tends to 0 as  $r$  tends to  $\infty$ .

If we know in advance that  $\lim_{|z| \rightarrow \infty} zf(z) = 0$ , then it would be sufficient to apply the lemma in Sect. 9.2.3. To prove that  $\int_{\Gamma(R)} f(z)e^{iz} dz$  tends to 0 with only the hypothesis of the theorem above, we use the following lemma:

♠ **Jordan Lemma:**

Let  $f(z)$  be a function defined in a sector of the half-plane  $y \geq 0$ . If  $\lim_{|z| \rightarrow \infty} f(z) = 0$ , the integral  $\int f(z)e^{iz} dz$  extended over the arc of the circle  $|z| = r$  contained in the sector tends to 0 as  $r$  tends to  $\infty$ .

*Proof* Let us put  $z = re^{i\theta}$  and let  $M(r)$  be the upper bound of  $|f(re^{i\theta})|$  as  $\theta$  varies, the point  $e^{i\theta}$  remaining in the sector. Then,

$$\left| \int f(z)e^{iz} dz \right| \leq M(r) \int_0^\pi e^{-r \sin \theta} r d\theta = 2M(r) \int_0^{\pi/2} e^{-r \sin \theta} r d\theta. \quad (9.24)$$

Since

$$\frac{\pi}{2} \leq \frac{\sin \theta}{\theta} \leq 1 \quad \text{for } 0 \leq \theta \leq \frac{\pi}{2},$$

we have

$$\int_0^{\pi/2} e^{-r \sin \theta} r \theta \leq \int_0^{\pi/2} e^{-\frac{2}{\pi} r \theta} r \theta \leq \int_0^{\infty} e^{-\frac{2}{\pi} r \theta} r \theta = \frac{\pi}{2}. \quad (9.25)$$

From (9.24) and (9.25), it follows that

$$\left| \int f(z) e^{iz} dz \right| \leq \pi M(r). \quad (9.26)$$

In view of our assumption  $\lim_{|z| \rightarrow \infty} f(z) = 0$ , the right-hand side of (9.26) vanishes as  $r \rightarrow \infty$ , which completes the proof. ♣

*Remark.*

1. If we have to calculate an integral

$$\int_{-\infty}^{\infty} f(x) e^{-ix} dx$$

that involves a *negative* imaginary exponential  $e^{-ix}$ , it would be necessary to integrate in the *lower* half-plane instead of the upper one because the function  $|e^{-iz}|$  is bounded in the lower half-plane  $y \leq 0$ . More generally, an integral of the form  $\int_{-\infty}^{\infty} f(x) e^{ax} dx$  (where  $a$  is complex constant) can be evaluated by integrating in the half-plane where  $|e^{az}| \leq 1$ .

2. Remember that  $\sin z$  and  $\cos z$  are not bounded in any half-plane. To evaluate integrals of the form

$$\int_{-\infty}^{\infty} f(x) \sin^n x dx \quad \text{and} \quad \int_{-\infty}^{\infty} f(x) \cos^n x dx,$$

we always express the trigonometric functions in terms of complex exponentials so that the preceding methods can be applied.

### 9.2.5 Type 4: Integrals of $f(x)/x^\alpha$

Consider integrals of the form

$$\int_0^{\infty} \frac{f(x)}{x^\alpha} dx,$$

where  $\alpha$  denotes a real number such that  $0 < \alpha < 1$ , and  $f(x)$  is a rational function with no pole on the positive real axis  $x \geq 0$ . In addition, we assume  $f(z)$  such that  $f(z)/z^{\alpha-1} \rightarrow 0$  in the limits  $z \rightarrow 0$  and  $z \rightarrow \infty$ .

To calculate such an integral, we consider the function

$$g(z) = \frac{f(z)}{z^\alpha}$$

of the complex variable  $z$ , defined in the plane with the positive real axis  $x \geq 0$  excluded. Let  $D$  be the open set thus defined. It is necessary to specify the branch of  $z^\alpha$  chosen in  $D$ , so we take the branch of the argument of  $z$  between 0 and  $2\pi$ . With this convention, we integrate  $g(z)$  along the closed path  $C(r, \varepsilon)$  as follows: we first trace the real axis from  $\varepsilon > 0$  to  $r > 0$ , then the circle  $\Gamma(r)$  of centered at the origin and radius  $r$  in the positive sense, then the real axis from  $r$  to  $\varepsilon$ , and finally, the circle  $\gamma(\varepsilon)$  of center 0 and radius  $\varepsilon$  in the negative sense. The integral

$$\int_{C(r, \varepsilon)} \frac{f(z)}{z^\alpha} dz$$

is equal to the sum of the residues of the poles of  $f(z)/z^\alpha$  contained in  $D$  if  $r$  has been chosen sufficiently large and  $\varepsilon$  sufficiently small. We have

$$\int_{C(r, \varepsilon)} \frac{f(z)}{z^\alpha} dz = \int_{\Gamma(r)} \frac{f(z)}{z^\alpha} dz + \int_{\gamma(\varepsilon)} \frac{f(z)}{z^\alpha} dz + (1 - e^{-2\pi i \alpha}) \int_\varepsilon^r \frac{f(x)}{x^\alpha} dx$$

because when the argument of  $z$  is equal to  $2\pi$ ,

$$z^\alpha = e^{2\pi i \alpha} |z|^\alpha.$$

From assumption,  $f(z)/z^{\alpha-1}$  tends to 0 when  $z$  tends to 0 or when  $|z|$  tends to infinity. Thus the integrals along  $\Gamma(r)$  and  $\gamma(\varepsilon)$  tend to 0 as  $r \rightarrow \infty$  and  $\varepsilon \rightarrow 0$ . On the limit, we have

$$(1 - e^{-2\pi i \alpha}) \int_0^\infty \frac{f(x)}{x^\alpha} dx = 2\pi i \sum \operatorname{Res} \left[ \frac{f(z)}{z^\alpha} \right]. \quad (9.27)$$

This relation allows us to calculate the original integral.

*Example* Try to evaluate the integral

$$I = \int_0^\infty \frac{dx}{x^\alpha(1+x)} \quad (0 < \alpha < 1).$$

Here we have

$$f(z) = \frac{1}{1+z},$$

where there is only one pole at  $z = -1$ . As the branch of the argument of  $z$  is equal to  $\pi$  at this point, the residue of  $f(z)/z^\alpha$  at this pole is equal to  $1/e^{\pi i \alpha}$ . Relation (9.27) then gives

$$I = \frac{\pi}{\sin \pi \alpha}.$$

### 9.2.6 Type 5: Integrals of $f(x) \log x$

The final type of integral to be noted is a class of the form

$$\int_0^\infty f(x) \log x dx,$$

where  $f$  is a rational function with no pole on the positive real axis  $x \geq 0$  and

$$\lim_{x \rightarrow \infty} x f(x) = 0.$$

This last condition ensures that the integral is convergent.

We consider the same open set  $D$  as for integrals of **Type 4** and the same path of integration. Here again, we must specify the branch chosen for  $\log z$ , and we choose the argument of  $z$  between 0 and  $2\pi$ . For a reason that will soon be apparent, we integrate the function  $f(z)(\log z)^2$  instead of  $f(z) \log z$ . Here again the integrals along the circles  $\Gamma(r)$  and  $\gamma(\varepsilon)$  tend to 0 as  $r \rightarrow \infty$  and  $\varepsilon \rightarrow 0$ , respectively.

When the argument  $z$  is equal to  $2\pi$ , we have

$$\log z = \log x + 2\pi i.$$

Thus we have the relation

$$\int_0^\infty f(x)(\log x)^2 dx - \int_0^\infty f(x)(\log x + 2\pi i)^2 dx = 2\pi i \sum \text{Res} [f(z)(\log z)^2].$$

and, hence,

$$-2 \int_0^\infty f(x) \log x dx - 2\pi i \int_0^\infty f(x) dx = \sum \text{Res} [f(z)(\log z)^2]. \quad (9.28)$$

By taking the imaginary part of the relation (9.28), we obtain the desired result:

$$\int_0^\infty f(x) \log x dx = -\frac{1}{2} \text{Im} \left\{ \sum \text{Res} [f(z)(\log z)^2] \right\}.$$

*Example* Consider the integral

$$I = \int_0^\infty \frac{\log x}{(1+x)^3} dx.$$

As the residue of  $(\log z)^2/(1+z)^3$  at the pole  $z = -1$  is equal to  $1 - i\pi$ , we find

$$I = -\frac{1}{2}.$$

## Exercises

1. Evaluate the integral defined by  $I = \int_0^{2\pi} \frac{d\theta}{(1 - a \cos \theta)^2}$  ( $0 < a < 1$ ).

**Solution:** Let  $z = e^{i\theta}$  and set  $C : |z| = 1$ . Then

$$I = \frac{4}{ia^2} \oint_C \frac{z dz}{[z^2 + (2z/a) + 1]^2}.$$

The integrand has two poles of second order at  $z = z_1, z_2$  ( $|z_1| < |z_2|$ ), which are the solutions of the equation  $g(z) = z^2 + (2z/a) + 1 = 0$ . Since  $0 < a < 1$ , only the pole  $z_1 = (-1 + \sqrt{1 - a^2})/a$  is found within  $C$ . The residue at  $z_1$  is given by

$$\begin{aligned} \text{Res}(z_1) &= \lim_{z \rightarrow z_1} \frac{d}{dz} \left[ (z - z_1)^2 \frac{z}{g(z)^2} \right] = \lim_{z \rightarrow z_1} \frac{d}{dz} \frac{z}{(z - z_2)^2} \\ &= -\frac{z_1 + z_2}{(z_1 - z_2)^3} = \frac{2/a}{(2\sqrt{1 - a^2}/a)^3}, \end{aligned}$$

and thus we obtain

$$I = \frac{4}{ia^2} \times 2\pi i \text{Res}(z_1) = \frac{2\pi}{(1 - a^2)^{3/2}}. \quad \clubsuit$$

2. Evaluate the integral  $I = \frac{1}{2\pi i} \oint_C \frac{e^z}{z^n} dz$  ( $C : |z| = 1$ ) for integer  $n$ .

**Solution:** For integers  $n \leq 0$ , it is apparent that  $I = 0$  since the integrand is analytic within and on  $C$ . For integers  $n > 0$ ,  $f(z) = e^z z^{-n}$  has a pole of order  $n$  at  $z = 0$ . Using the residue theorem, we have  $I = 1/(n-1)!$ .  $\clubsuit$

3. Calculate the integral  $I = \int_{-\infty}^{\infty} \frac{dx}{(1 + x^2)^{n+1}}$ .

**Solution:** Define the function  $f(z) = 1/(1 + z^2)^{n+1}$ , and set the semicircle  $C$  as shown in Fig. 9.1. Within  $C$ ,  $f(z)$  has the pole of  $(n+1)$ th order at  $z = i$ , and its residue reads

$$\begin{aligned} \frac{1}{n!} \left[ \frac{d^n}{dz^n} \frac{(z - i)^{n+1}}{(1 + z^2)^{n+1}} \right]_{z=i} &= \frac{1}{n!} \left[ \frac{d^n}{dz^n} (z + i)^{-(n+1)} \right]_{z=i} \\ &= \frac{(-1)^n (n+1)(n+2) \cdots 2n}{n!} (2i)^{-(2n+1)} \\ &= \frac{(2n)!}{2^{2n} (n!)^2} \frac{1}{2i}. \end{aligned}$$

Hence, in view of Cauchy's theorem, we have

$$\oint_C f(z)dz = \frac{\pi(2n)!}{2^{2n}(n!)^2}. \quad (9.29)$$

We now observe that

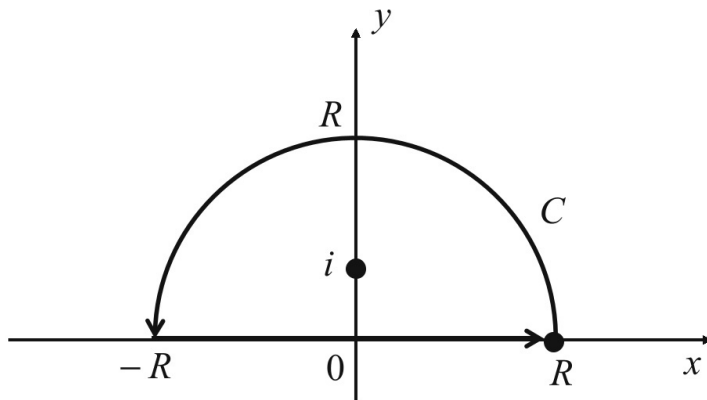
$$\oint_C f(z)dz = \int_{-R}^R \frac{dx}{(1+x^2)^{n+1}} + \int_{\Gamma} \frac{dz}{(1+z^2)^{n+1}}, \quad (9.30)$$

where  $\Gamma$  denotes the upper half-circle. Since  $|1+z^2| \geq R^2 - 1$  on  $C$ , the second integral in the limit  $R \rightarrow \infty$  yields

$$\left| \int_{\Gamma} \frac{dz}{(1+z^2)^{n+1}} \right| \leq \frac{\pi R}{(R^2 - 1)^{n+1}} \rightarrow 0. \quad (9.31)$$

From (9.29)–(9.31), we conclude that

$$I = \frac{\pi(2n)!}{2^{2n}(n!)^2}. \quad \clubsuit$$



**Fig. 9.1.** The integration path used in Exercise 3

4. Calculate the integral  $I = \int_0^\pi \log(1 - 2r \cos \theta + r^2) d\theta$ , where  $r \neq 1$ .

**Solution:** First we assume that  $r < 1$ . Observe that the function  $\log(1-z)/z = -1 - (z/2) - (z^2/3) - \cdots$  is analytic for  $|z| \leq r < 1$ . Hence, if we set the circle  $C : |z| = r$ , we have

$$\oint_C \frac{\log(1-z)}{z} dz = i \int_0^{2\pi} \log(1-z) d\theta = 0. \quad (9.32)$$

Since  $|1-z|^2 = 1 - 2r \cos \theta + r^2$  on  $C$ , the real component of the second integral in (9.32) reads  $(i/2) \int_0^{2\pi} \log(1 - 2r \cos \theta + r^2) d\theta = 0$ , so we get

$$I = 0 \quad \text{for } r < 1.$$

Next we consider the case of  $r > 1$ . Set  $s = 1/r < 1$  to obtain

$$\begin{aligned} 0 &= \int_0^\pi \log(1 - 2s \cos \theta + s^2) d\theta = \int_0^\pi \log\left(1 - \frac{2}{r} \cos \theta + \frac{1}{r^2}\right) d\theta \\ &= \int_0^\pi [\log(1 - 2r \cos \theta + r^2) - \log r^2] d\theta. \end{aligned}$$

Hence, we conclude that

$$I = 2\pi \log r \quad \text{for } r > 1. \quad \clubsuit$$

5. Calculate the integral  $I = \int_0^\infty \frac{x^{\alpha-1}}{1+x} dx$ , where  $0 < \alpha < 1$ .

**Solution:** Consider the power function

$$z^\beta = e^{\beta \log z} = e^{\beta(\log |z| + i \arg z)}$$

with  $-1 < \beta < 0$ . Its branch for  $0 < \arg z < 2\pi$  is single-valued on the domain  $D$  enclosed by the contour  $C = AB + \Gamma + B'A' + \gamma$  depicted in Fig. 9.2. Let the radius  $r$  of the circle  $\gamma$  be sufficiently small and that  $R$  of  $\Gamma$  be sufficiently large. Then, the pole  $z = -1$  of the function  $f(z) = z^\beta/(1+z)$  is located within  $C$  so that we have

$$\oint_C f(z) dz = \int_r^R \frac{x^\beta}{1+x} + \int_\Gamma f(z) dz - \int_r^R \frac{x^\beta e^{\beta \cdot 2\pi i}}{1+x} + \int_\gamma f(z) dz. \quad (9.33)$$

Observe that

$$\left| \int_\Gamma f(z) dz \right| \leq \int_\Gamma \frac{|z^\beta|}{|1+z|} |dz| \leq \frac{2\pi R^{\beta+1}}{R-1} \rightarrow 0 \quad (R \rightarrow \infty)$$

and

$$\left| \int_\gamma f(z) dz \right| \leq \frac{2\pi r^{\beta+1}}{1-r} \rightarrow 0 \quad (r \rightarrow 0).$$

Take the limits  $R \rightarrow \infty$  and  $r \rightarrow 0$  on both sides of (9.33) to yield

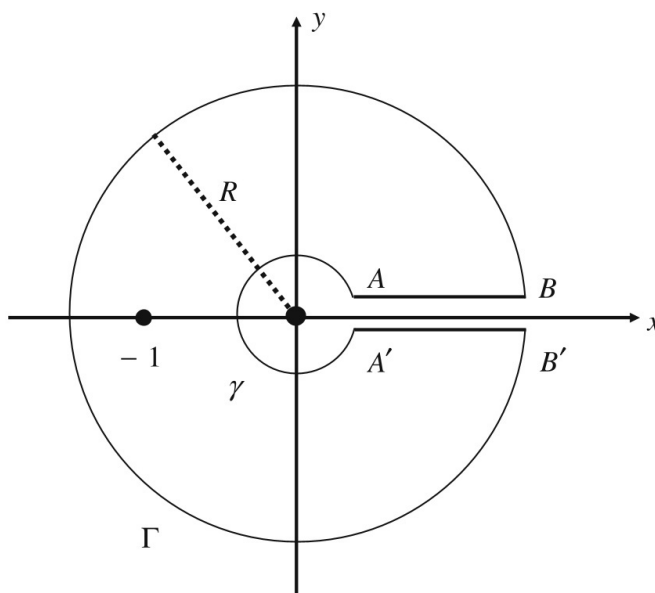
$$(1 - e^{\beta \cdot 2\pi i}) \int_0^\infty \frac{x^\beta}{1+x} dx = \text{Res} \left[ \frac{z^\beta}{1+z}, -1 \right] = 2\pi i \lim_{z \rightarrow -1} z^\beta = 2\pi i e^{\beta \cdot \pi i},$$

which then gives us

$$\int_0^\infty \frac{x^\beta}{1+x} dx = 2\pi i \frac{e^{\beta \cdot \pi i}}{1 - e^{\beta \cdot 2\pi i}}.$$

Since  $\beta = \alpha - 1$ , the above result is equivalent to

$$I = \frac{\pi}{\sin \alpha \pi}. \quad \clubsuit$$



**Fig. 9.2.** Integration path  $C = AB + \Gamma + B'A' + \gamma$  used in Exercise 5

## 9.3 More Applications of Residue Calculus

### 9.3.1 Integrals on Rectangular Contours

The integrals discussed so far are evaluated using the residue theorem based on a circular (or semicircular) contour whose radius is eventually made to be infinitely large or infinitely small. However, there are other integrals that can be evaluated by the residue theorem that do not have to be closed with a circle. Several examples are given below.

Let us consider the integral

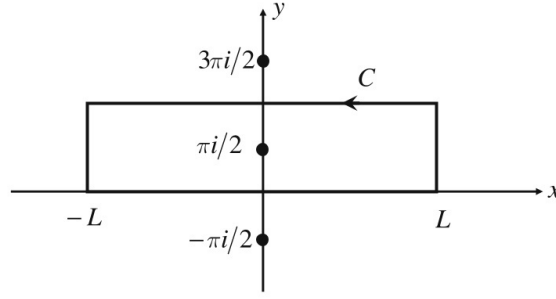
$$I = \int_{-\infty}^{\infty} \frac{xe^x}{(1+e^{2x})^2} dx.$$

To evaluate it, we examine the contour integral

$$J = \oint_C \frac{ze^z}{(1+e^{2z})^2} dz \quad (9.34)$$

around the rectangular contour shown in Fig. 9.3. Beginning at the lower left hand corner of the rectangle,

$$\begin{aligned} J = & \int_{-L}^L \frac{xe^x}{(1+e^{2x})^2} dx + \int_0^\pi \frac{(L+iy)e^{L+iy}}{(1+e^{2(L+iy)})^2} idy + \int_L^{-L} \frac{(x+i\pi)e^{x+i\pi}}{(1+e^{2(x+i\pi)})^2} dx \\ & + \int_\pi^0 \frac{(-L+iy)e^{-L+iy}}{(1+e^{2(-L+iy)})^2} idy. \end{aligned} \quad (9.35)$$



**Fig. 9.3.** Rectangular contour surrounding the path  $z = \pi i/2$

In the limit  $L \rightarrow \infty$ , the second and fourth integral of (9.35) go to zero, since in this limit the magnitude of  $e^{2(L+iy)}$  and  $e^{2(-L+iy)}$  become very large or very small, respectively, compared to unity. Hence, we have

$$\begin{aligned} \lim_{L \rightarrow \infty} J &= \int_{-\infty}^{\infty} \frac{x e^x}{(1 + e^{2x})^2} dx + \int_{\infty}^{-\infty} \frac{(x + i\pi) e^{x+i\pi}}{(1 + e^{2(x+i\pi)})^2} dx, \\ &= I + \int_{-\infty}^{\infty} \frac{(x + i\pi) e^{x+i\pi}}{(1 + e^{2(x+i\pi)})^2} dx = 2I - i\pi \int_{-\infty}^{\infty} \frac{e^x}{(1 + e^{2x})^2} dx, \end{aligned} \quad (9.36)$$

where we have used the expressions  $e^{x+i\pi} = -e^x$  and  $e^{2(x+i\pi)} = e^{2x}$ . As a result, the integral  $I$  to be evaluated is expressed in terms of  $J$  as

$$I = \frac{1}{2} \lim_{L \rightarrow \infty} J + \frac{i\pi}{2} \int_{-\infty}^{\infty} \frac{e^x}{(1 + e^{2x})^2} dx. \quad (9.37)$$

The contour integral  $J$  is readily evaluated by employing the residue theorem. Looking back to the definition (9.34), we see that  $J$  has second-order poles at the values of  $z$  for which  $e^{2z} = -1$ . These values are

$$z = \pm \frac{i\pi}{2}, \pm \frac{3i\pi}{2}, \dots, \pm i \left( N + \frac{1}{2} \right) \pi,$$

where  $N$  is a nonnegative integer. Note that only the pole at  $z = i\pi/2$  is enclosed in the rectangle (see Fig. 9.3). Hence, using the ratio method (see Sect. 9.1.4) we have

$$J = 2\pi i \operatorname{Res} \left( \frac{i\pi}{2} \right) = 2\pi i \cdot 2 \frac{p'(i\pi/2)}{q''(i\pi/2)} = \frac{-\pi(2 + i\pi)}{4}, \quad (9.38)$$

where  $p(z) = ze^z$  and  $q(z) = (1 + e^{2z})^2$  are constituents of the integrand in (9.34).

The latter integral of (9.37) is evaluated by substituting  $w = e^x$ , and it follows that

$$\int_{-\infty}^{\infty} \frac{e^x}{(1 + e^{2x})^2} dx = \int_0^{\infty} \frac{dw}{(1 + w^2)^2} = \frac{1}{2} \int_{-\infty}^{\infty} \frac{dw}{(1 + w^2)^2}, \quad (9.39)$$

Thus, applying the residue theorem yields

$$\frac{1}{2} \int_{-\infty}^{\infty} \frac{dw}{(1+w^2)^2} = \frac{1}{2} \cdot 2\pi i \cdot \text{Res}(i) = \pi i \lim_{w \rightarrow i} \frac{d^2}{dw^2} \frac{1}{(w+i)^2} = \frac{\pi}{4}. \quad (9.40)$$

From (9.38) and (9.40), we finally obtain

$$I = -\frac{\pi}{4}(2 + i\pi) + \frac{\pi}{4} = -\frac{\pi}{4}(1 + i\pi).$$

### 9.3.2 Fresnel Integrals

We would like to derive the equations

$$\int_0^{\infty} \cos(kx^2) dx = \int_0^{\infty} \sin(kx^2) dx = \frac{1}{2} \sqrt{\frac{\pi}{2k}}$$

with a real positive constant  $k$ . These are known as the **Fresnel cosine integral** and **Fresnel sine integral**. Integrals of this type are encountered in the study of a phenomenon called diffraction, which is exhibited by all types of waves such as light and sound.

In this connection we consider the integral

$$I = \oint_C e^{ikz^2} dz \quad (k > 0) \quad (9.41)$$

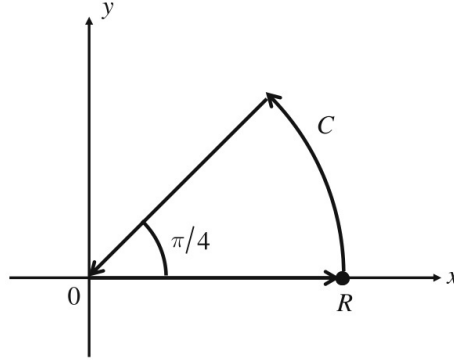
around the contour shown in Fig. 9.4. The integral variable  $z$  becomes  $z = x$  on the segment along the real axis,  $z = Re^{i\phi}$  ( $0 \leq \phi \leq \pi/4$ ) along the large (ultimately infinite) arc, and  $z = x(1+i)$  along the slanted segment defined by  $y = x$ . Therefore, with  $(1+i)^2 = 2i$ , we have

$$\begin{aligned} \lim_{R \rightarrow \infty} \oint_C e^{ikz^2} dz &= \int_0^{\infty} e^{ikx^2} dx + \lim_{R \rightarrow \infty} \left[ \int_0^{\pi/4} e^{ikR^2 e^{2i\phi}} iR d\phi \right] \\ &\quad + (1+i) \int_{\infty}^0 e^{-2kx^2} dx. \end{aligned} \quad (9.42)$$

Our objective is to evaluate the real and imaginary parts of the first integral on the right-hand side of (9.42). Then, evaluations of the other integrals shown in (9.42) complete the computation.

First, we readily obtain

$$\oint_C e^{ikz^2} dz = 0, \quad (9.43)$$



**Fig. 9.4.** Contour for evaluating the integral (9.41)

since there are no poles within the contour of Fig. 9.4.

Second, we consider the integral along the arc, which is given in the second term on the right-hand side of (9.42). On the large arc, we have

$$\left| Re^{ikR^2 e^{2i\phi}} \right| = \left| Re^{ikR^2 \cos(2\phi)} e^{-kR^2 \sin(2\phi)} \right| \leq Re^{-kR^2 \sin(2\phi)},$$

where the sign of  $\sin(2\phi)$  is always nonnegative in the range  $0 \leq \phi \leq \pi/4$ . Hence,

$$\lim_{R \rightarrow \infty} Re^{-kR^2 \sin(2\phi)} = 0, \quad (9.44)$$

so that the integral along the arc vanishes in the limit  $R \rightarrow \infty$ . In fact, l'Hôpital's rule states that for  $a \geq 0$ ,

$$\lim_{R \rightarrow \infty} \frac{R}{e^{aR^2}} = \lim_{R \rightarrow \infty} \frac{1}{2aRe^{aR^2}} = 0.$$

Finally we examine the integral along the slanted segment, i.e., the third term on the right-hand side of (9.42). To evaluate it, we consider the quantity

$$\begin{aligned} J &\equiv \left( \int_{-\infty}^{\infty} e^{-2kx^2} dx \right)^2 = \int_{-\infty}^{\infty} e^{-2kx^2} dx \cdot \int_{-\infty}^{\infty} e^{-2ky^2} dy \\ &= \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy e^{-2k(x^2+y^2)}. \end{aligned}$$

In terms of the polar coordinates, it yields

$$J = \int_0^{\infty} dr \int_0^{2\pi} d\theta r e^{-2kr^2} = \int_0^{\infty} \frac{d(r^2)}{2} \int_0^{2\pi} d\theta e^{-2kr^2} = \frac{\pi}{2k}.$$

and we have the Gaussian integral given by

$$\int_{-\infty}^{\infty} e^{-2kx^2} dx = \sqrt{\frac{\pi}{2k}},$$

so that

$$(1+i) \int_{-\infty}^0 e^{-2kx^2} dx = \frac{1+i}{2} \sqrt{\frac{\pi}{8k}}. \quad (9.45)$$

Substituting the results of (9.43), (9.44), and (9.45) into (9.42), we find that

$$\int_0^{\infty} e^{ikx^2} dx = \frac{1+i}{2} \sqrt{\frac{\pi}{2k}}. \quad (9.46)$$

Writing the exponential in trigonometric form and equating the real and imaginary parts of both sides of (9.42), we obtain the Fresnel integral:

$$\int_0^{\infty} \cos(kx^2) dx = \int_0^{\infty} \sin(kx^2) dx = \frac{1}{2} \sqrt{\frac{\pi}{2k}}. \clubsuit$$

### 9.3.3 Summation of Series

Our final application of the residue theorem is the summation of a series  $\sum_{n=-\infty}^{\infty} f(n)$ . Using this method, we can convert a certain type of series to simple forms such as

$$\sum_{n=-\infty}^{\infty} \frac{1}{(a+n)^2} = \frac{\pi^2}{\sin^2(\pi a)} \quad (9.47)$$

and

$$\sum_{n=1}^{\infty} \frac{2x}{x^2 + n^2\pi^2} = \coth x - \frac{1}{x}.$$

This technique is particularly useful, for instance, to express a power series solution of a differential equation in a simple closed form. In fact, this device is generalized for various series summations as shown below.

#### ♠ Theorem:

An infinite series of functions  $f(n)$  with respect to an integer  $n$  is given by

$$\sum_{n=-\infty}^{\infty} f(n) = - \sum_{n=-\infty}^{\infty} \text{Res}(g, a_n), \quad (9.48)$$

where  $\text{Res}(g, a_n)$  is the residue of the specific function

$$g(z) = \frac{\pi f(z)}{\tan(\pi z)}$$

at the  $n$ th pole of  $f(z)$  located at  $z = a_n$ .

According to this theorem, we see that if the number of poles of  $f(z)$  is finite and the values of  $\text{Res}(g, a_n)$  are readily obtained, the series on the left-hand side of (9.48) is written in a simple form.

*Proof* The key point is to use a function given by  $\pi/\tan(\pi z)$ . This function has simple poles at  $z = 0, \pm 1, \pm 2, \dots$ , each with residue 1 evaluated as

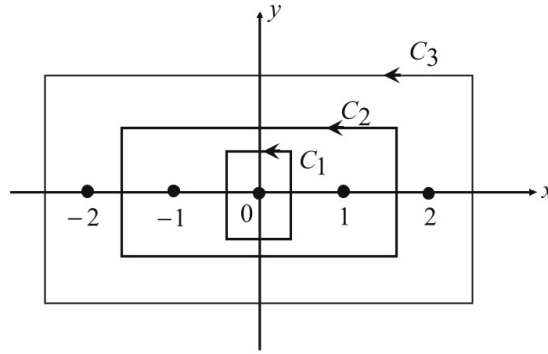
$$\lim_{z \rightarrow n} \frac{\pi}{\tan(\pi z)} \cdot (z - n) = \lim_{z \rightarrow n} \frac{\pi}{\pi / \cos^2(\pi z)} = 1,$$

where we used **l'Hôpital's rule** (see Exercise 3 in Sect. 8.1). In addition, the function  $\pi/\tan(\pi z)$  is bounded at infinity except on the real axis. To derive (9.48), let us consider the contour integral

$$\oint_{C_1} \frac{\pi f(z)}{\tan(\pi z)} dz \quad (9.49)$$

around the contour  $C_1$  shown in Fig. 9.5. Here  $f(z)$  is assumed to have no branch points or essential singularities anywhere. Since only the pole at  $z = 0$  is found within  $C_1$ , the contour integral equals  $2\pi i$  times the residue of the integrand at  $z = 0$ , which is  $f(0)$ , i.e.,

$$\oint_{C_1} \frac{\pi f(z)}{\tan(\pi z)} dz = 2\pi i f(0).$$



**Fig. 9.5.** A sequence of rectangular contours to derive equation (9.48)

Next, the integral around contour  $C_2$  is

$$\oint_{C_2} \frac{\pi f(z)}{\tan(\pi z)} dz = 2\pi i [f(0) + f(1) + f(-1) + \text{Res}(g, a_1)],$$

where  $\text{Res}(g, a_1)$  stems from the contribution of the pole of  $f(z)$  located at  $z = a_1$ . Finally, for a contour at infinity, the integral must be

$$\oint_{C_\infty} \frac{\pi f(z)}{\tan(\pi z)} dz = 2\pi i \left\{ \sum_{n=-\infty}^{\infty} [f(n) + \text{Res}(g, a_n)] \right\}. \quad (9.50)$$

If  $|zf(z)| \rightarrow 0$  as  $|z| \rightarrow \infty$ , the infinite contour integral is zero so that we successfully obtain the equation:

$$\sum_{n=-\infty}^{\infty} f(n) = - \sum_{n=-\infty}^{\infty} \text{Res}(g, a_n). \quad \clubsuit \quad (9.51)$$

### 9.3.4 Langevin and Riemann zeta Functions

Our present aim is to establish the equivalence between **Langevin's function**,

$$\coth x - (1/x),$$

and the sum

$$\sum_{n=1}^{\infty} \frac{2x}{x^2 + n^2\pi^2}.$$

Letting  $f(z) = 2x/(x^2 + z^2\pi^2)$ , and using the above equation, we obtain

$$\sum_{m=-N}^N \frac{2x}{x^2 + n^2\pi^2} = \frac{1}{2\pi i} \oint_C \pi \cot \pi z f(z) dz - \sum_{\text{poles}} \text{Res} [\pi \cot(\pi z) f(z)],$$

where  $C$  is a closed contour, say, a rectangle, enclosing the points  $z = 0, \pm 1, \dots$ . Now let the length and width of the rectangle  $C$  approach  $\infty$ . As this happens,

$$\left| \frac{1}{2\pi i} \oint_C \pi \cot \pi z f(z) dz \right| \leq \frac{1}{2} \oint_C \pi |\cot \pi z| \left| \frac{2x}{x^2 + n^2\pi^2} \right| |dz| \rightarrow 0. \quad (9.52)$$

Hence, we have

$$\begin{aligned} \sum_{m=-\infty}^{\infty} \frac{2x}{x^2 + n^2\pi^2} &= -\text{Res} \left[ \frac{(\pi \cot \pi z) 2x}{x^2 + z^2\pi^2} \right]_{z=\pm ix/\pi} \\ &= -\frac{2x}{\pi} \left[ \frac{\cot(ix)}{2ix/\pi} + \frac{\cot(-ix)}{-2ix/\pi} \right] \\ &= 2i \cot(ix) = 2 \coth x. \end{aligned}$$

This result can be rewritten as

$$2 \sum_{m=1}^{\infty} \frac{2x}{x^2 + n^2\pi^2} + \frac{2}{x} = 2 \coth x$$

or

$$\coth x - \frac{1}{x} = \sum_{m=1}^{\infty} \frac{2x}{x^2 + n^2\pi^2}, \quad (9.53)$$

which establishes the result we stated at the outset.

*Remark.* To see that the integral in (9.52) vanishes as  $z \rightarrow \infty$ , we observe that

$$|\cot \pi z| = \frac{|\cos \pi z|}{|\sin \pi z|} = \sqrt{\frac{\cos^2 \pi x + \sinh^2 \pi y}{\sin^2 \pi x + \sinh^2 \pi y}}.$$

If we choose the rectangle whose vertical sides cross the  $x$ -axis at a large enough half-integer, say,  $x = 10^5 + \frac{1}{2}$  so that  $\cos \pi x = 0$  and  $\sin \pi x = 1$ , then over these sides of the rectangle

$$|\cot \pi z| = \sqrt{\frac{\sinh^2 \pi y}{1 + \sinh^2 \pi y}} = |\tanh \pi y| \leq 1.$$

Over the horizontal sides of the rectangle,  $\lim_{z \rightarrow \infty} |\cot \pi z| = 1$ . Thus the integrand goes as  $|1/z^2|$  as  $|z| \rightarrow \infty$ , and the integral vanishes.

If we integrate both sides of (9.53) from 0 to  $x$ , we get

$$\sum_{m=1}^{\infty} \ln \left( 1 + \frac{x^2}{m^2 \pi^2} \right) = \ln \left[ \prod_{m=1}^{\infty} \left( 1 + \frac{x^2}{m^2 \pi^2} \right) \right] = \ln \left( \frac{\sinh x}{x} \right).$$

Hence,

$$\frac{\sinh x}{x} = \prod_{m=1}^{\infty} \left( 1 + \frac{x^2}{m^2 \pi^2} \right).$$

We may extend this result to all  $z$  in the complex plane by **analytic continuation**. Then setting  $x = i\theta$  with  $\theta$  real, we obtain

$$\sin \theta = \theta \prod_{n=1}^{\infty} \left( 1 - \frac{\theta^2}{n^2 \pi^2} \right).$$

This infinite product formula displays all the zeros of  $\sin \theta$  explicitly. It represents the complete factorization of the Taylor series and can, in fact, be taken as the definition of the sine function.

By equating coefficients of the  $\theta^3$  term of both sides of the above equation, we obtain a useful sum:

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6},$$

which is a special value of the **Riemann zeta function**,

$$\zeta(z) = \sum_{n=1}^{\infty} \frac{1}{n^z}.$$

## Exercises

1. Evaluate  $\sum_{n=-\infty}^{\infty} \frac{1}{(a+n)^2}$  by considering the contour integral:

$$I = \oint_C \frac{\pi}{\tan(\pi z)} \frac{1}{(a+z)^2} dz,$$

where  $a$  is not an integer and  $C$  is a circle of large radius,

**Solution:** In order to use equation (9.48), we define

$$f(z) = \frac{1}{(a+z)^2} \quad \text{and} \quad g(z) = \frac{\pi}{\tan(\pi z)} \cdot \frac{1}{(a+z)^2}.$$

Since the integrand  $g(z)$  has simple poles at  $z = 0, \pm 1, \pm 2, \dots$  and a double pole at  $z = -a$ , evaluation of  $\text{Res}(g, -a)$  completes the problem [see (9.48)]. To find the residue at  $z = -a$ , set  $z = -a + \xi$  for small  $\xi$  and determine the coefficient of  $\xi^{-1}$ :

$$\begin{aligned} \frac{\pi}{\tan(\pi z)} \frac{1}{(a+z)^2} &= \frac{\pi}{\xi^2} \frac{1}{\tan(-a\pi + \xi\pi)} \\ &= \frac{\pi}{\xi^2} \left\{ \frac{1}{\tan(-a\pi)} + \xi \left[ \frac{d}{dz} \frac{1}{\tan(\pi z)} \right]_{z=-a} + \dots \right\}. \end{aligned}$$

It follows from (9.54) that the residue at the double pole  $z = -a$  is

$$\pi \left[ \frac{d}{dz} \frac{1}{\tan(\pi z)} \right]_{z=-a} = \pi \left[ \frac{-\pi}{\sin^2(\pi z)} \right]_{z=-a} = -\frac{\pi^2}{\sin^2(\pi a)}.$$

Therefore, it is readily seen from (9.48) that

$$\sum_{n=-\infty}^{\infty} \frac{1}{(a+n)^2} = \frac{\pi^2}{\sin^2 \pi a}. \quad \clubsuit$$

## 9.4 Argument Principle

### 9.4.1 The Principle

It may occur that a function  $f(z)$  has several zeros and poles simultaneously in a domain  $D$ . If we denote the number of such zeros and poles by  $N_0$  and  $N_\infty$ , respectively, these numbers are related to one another as stated below.

♠ **Argument principle:**

Let  $f(z)$  be an analytic function within a closed contour  $C$  except at a finite number of poles. If  $f(z) \neq 0$  on  $C$ , then

$$\frac{1}{2\pi i} \oint_C \frac{f'(z)}{f(z)} dz = N_0 - N_\infty, \quad (9.54)$$

where  $N_0$  and  $N_\infty$  are the numbers of zeros and poles of  $f(z)$  in  $C$ , respectively. Both zeros and poles are to be counted with their multiplicities.

*Proof* By the residue theorem, the integral

$$\frac{1}{2\pi i} \oint_C \frac{f'(z)}{f(z)} dz$$

is equal to the sum of the residues of the logarithmic derivative of  $f(z)$  in  $D$ , i.e.,

$$g(z) = \frac{f'(z)}{f(z)} = \frac{d[\log f(z)]}{dz}.$$

The only possible singularities of  $g(z)$  in  $D$  coincide with the zeros and poles of  $f(z)$ . In order to determine the residue of  $g(z)$  at a zero of  $f(z)$ , we observe that in the neighborhood of a zero  $a$  of the  $n$ th order,  $f(z)$  has an expansion

$$f(z) = (z - a)^n [c_1 + c_2(z - a) + \cdots], \quad c_1 \neq 0.$$

We therefore have

$$f(z) = (z - a)^n f_1(z),$$

where  $f_1(z) \neq 0$  in a certain neighborhood of  $z = a$ . Hence,

$$\log f(z) = n \log(z - a) + \log f_1(z),$$

and

$$\frac{f'(z)}{f(z)} = \frac{n}{z - a} + \frac{f_1'(z)}{f_1(z)},$$

where the last term is analytic at  $z = a$ . It follows that the residue of  $g(z)$ , which is called the **logarithmic residue** of  $f(z)$  at  $z = a$  is  $n$ , i.e., it is equal to the order of the zero of  $f(z)$  at  $z = a$ . If the zeros of  $f(z)$  in  $D$  are counted with their multiplicities, the sum of the logarithmic residues of  $f(z)$  at the zeros of  $f(z)$  in  $D$  will be equal to the number of zeros.

We now turn to the poles of  $f(z)$  in  $D$ . If  $z = b$  is a pole of order  $m$ , we have near it an expansion

$$\begin{aligned}
f'(z) &= \frac{c_1}{(z-b)^m} + \cdots + \frac{c_m}{z-b} + c_{m+1} + \cdots \\
&= \frac{1}{(z-b)^m} [c_1 + c_2(z-b) + \cdots] \\
&= \frac{f_2(z)}{(z-b)^m},
\end{aligned}$$

where  $f_2(z)$  is analytic at  $z = b$  and  $f_2(z) \neq 0$ . Hence,

$$\frac{f'(z)}{f(z)} = -\frac{m}{z-b} + \frac{f_2'(z)}{f_2(z)},$$

which shows that the logarithmic residue of  $f(z)$  at a pole of  $f(z)$  of order  $m$  is  $-m$ . If the poles of  $f(z)$  in  $D$  are counted with their multiplicities, the sum of the logarithmic residues of  $f(z)$  at the points of  $f(z)$  in  $D$  will be equal to minus the number of these poles. Since  $g(z)$  has no singularities in  $D$  except at the zeros and poles of  $f(z)$ , we have proven our theorem. ♣

*Remark.* If we replace  $f(z)$  in (9.54) by  $f(z) - a$ , this formula will yield the difference between the number of zeros and the poles of  $f(z) - a$ . Since the latter are identical with the poles of  $f(z)$ , we find that

$$\frac{1}{2\pi i} \oint_C \frac{f'(z)}{f(z) - a} dz = N_a - N_\infty,$$

where  $N_a$  indicates how often the value of  $a$  is taken by  $f(z)$  in  $D$ .

*Examples* **1.** For  $f(z) = z^2$  and  $C : |z| = 1$ ,  $N_0 = 2$  and  $N_\infty = 0$  so that we have

$$\frac{1}{2\pi i} \oint_C \frac{f'(z)}{f(z)} dz = 2.$$

In fact, the integral reads

$$\frac{1}{2\pi i} \oint_C \frac{f'(z)}{f(z)} dz = \frac{1}{2\pi i} \oint_C \frac{2z}{z^2} dz = \frac{1}{2\pi i} \times 2 \times 2\pi i = 2.$$

**2.** For  $f(z) = z/(z-a)$  and  $C : |z| = R$ ,  $N_0 = 1$  and

$$N_\infty = \begin{cases} 1 & \text{if } R > a, \\ 0 & \text{if } R < a. \end{cases}$$

Hence, we have

$$\frac{1}{2\pi i} \oint_C \frac{f'(z)}{f(z)} dz = \begin{cases} 0 & \text{if } R > a, \\ 1 & \text{if } R < a. \end{cases} \quad (9.55)$$

Indeed,  $f(z) = 1 + [a/(z-a)]$ ,  $f'(z) = -1/(z-a)^2$ ,  $f'/f = (1/z) - [1/(z-a)]$ , which yields (9.55).

### 9.4.2 Variation of the Argument

Equation (9.54) can be brought into a different form in which its geometric character becomes more apparent. If we write

$$\varphi = \arg f(z), \quad f(z) = |f(z)|e^{i\varphi},$$

we obtain

$$\begin{aligned} \frac{1}{2\pi i} \oint_C \frac{f'(z)}{f(z)} dz &= \frac{1}{2\pi i} \oint_C d \log f(z) \\ &= \frac{1}{2\pi i} \oint_C [d \log |f(z)| + i d\varphi] \\ &= \frac{1}{2\pi i} \oint_C d \log |f(z)| + \frac{1}{2\pi} \oint_C d\varphi. \end{aligned}$$

Recall that  $\log w(z)$  is a many-valued function of  $w$ . If  $\log w$  is continued along a closed curve that surrounds to origin, we shall not return to the value of  $\log w$  with which we started. However, this many-valuedness is confined to  $\text{Im}(\log w) = \arg w$ , i.e.,  $\text{Re}(\log w) = \log |w|$  is single-valued. If we write  $w = f(z)$ , it follows that

$$\oint_C d \log |f(z)| = 0.$$

In fact,

$$\int_{z_1}^{z_2} d \log |f(z)| = \log |f(z_2)| - \log |f(z_1)|,$$

and if the integration is performed over a closed contour, the terminals  $z_1$  and  $z_2$  of the integration coincide; moreover, owing the single-valuedness of  $\log |f(z)|$ , the value of the integral is zero. Hence, we have

$$\frac{1}{2\pi i} \oint_C \frac{f'(z)}{f(z)} dz = \frac{1}{2\pi} \oint_C d\varphi, \quad (9.56)$$

where  $\varphi = \arg f(z)$ .

To interpret (9.56), we observe that

$$\int_{z_1}^{z_2} d\varphi = \varphi(z_2) - \varphi(z_1) = \arg f(z_2) - \arg f(z_1)$$

is the quantitative change in the argument of  $f(z)$ , which is called the **variation of the argument** of  $f(z)$ . The integral  $\oint_C d\varphi$  is therefore the total variation of  $\arg f(z)$  if  $z$  describes the entire boundary  $C$  of the domain

$D$ . It is clear that the value of this integral must be an integral multiple of  $2\pi$ . If  $z$  describes  $C$ , the point  $f(z)$  describes a closed curve  $C'$ , and if  $C'$  surrounds the origin  $m$  times in the positive (counterclockwise) direction, the increase in  $\arg f(z)$  along  $C'$  will be  $2m\pi$ . In view of (9.54) and (9.56), we obtain the theorem below.

♠ **Theorem:**

Let the domain  $D$  be bounded by one or more closed contours  $C$  and let a function  $f(z)$  be single-valued and analytic apart from a finite number of poles. If  $N_0$  and  $N_\infty$  denote the number of zeros and poles of  $f(z)$  in  $D$ , respectively, and  $f(z) \neq 0$  on  $C$ , then

$$\frac{1}{2\pi} \Delta_c = N_0 - N_\infty,$$

where  $\Delta_c$  denotes the **total variation** of  $\arg f(z)$ .

### 9.4.3 Extentson of the Argument Principle

The argument principle can be extended to the case in which  $f(z)$  has zeros or poles on the boundary  $C$  of the domain  $D$ . Suppose that  $f(z_0) = 0$ , where  $z_0$  is situated on  $C$ . Let  $f(z)$  be analytic at  $z_0$ ; then we have

$$f(z) = (z - z_0)^m f_1(z), \quad f_1(z_0) \neq 0,$$

if  $m$  is the multiplicity of the zero. In view of the relation

$$\log f(z) = m \log(z - z_0) + \log f_1(z),$$

it follows that

$$\arg f(z) = m \arg(z - z_0) + \arg f_1(z).$$

At  $z = z_0$ ,  $f_1(z) \neq 0$  and  $\log f_1(z)$  is analytic. Hence,  $\arg f_1(z)$  will vary continuously if  $z$  varies along  $C$  and passes through  $z = z_0$ , but the expression  $\arg(z - z_0)$  shows a different behavior. Since this is the angle between the parallel to the positive axis through  $z_0$  and the linear segment drawn from  $z_0$  to  $z$ , it is clear that if  $z_0$  is passed  $\arg(z - z_0)$  jumps by the amount  $\pi$ . The contribution of this zero to  $\arg f(z)$  will be  $m\pi$ , i.e., one-half of what it would have been if the zero were situated in the interior of  $D$ . If  $z = z_0$  is a pole of order  $m$ , its contribution to  $\arg f(z)$  will be  $-m\pi$ . This follows immediately from the fact that  $f(z)^{-1}$  has a zero of order  $m$  at  $z_0$  and that

$$\log[f(z)^{-1}] = -\log f(z).$$

We therefore have the following extension of the argument principle.

♠ **Extended argument principle:**

The argument principle remains valid if  $f(z)$  has poles and zeros on the boundary, provided that these poles and zeros are counted with half their multiplicities.

#### 9.4.4 Rouché Theorem

As an application of the argument principle, we prove the following result, known as the **Rouché theorem**.

♠ **Rouché theorem:**

If the function  $f(z)$  and  $g(z)$  are analytic and single-valued in a domain  $D$  and on its boundary  $C$  and if  $|g(z)| < |f(z)|$  on  $C$ , then the number of zeros of the function  $f(z) + g(z)$  within  $D$  is equal to that of zeros of  $f(z)$ .

*Proof* We have

$$\log [f(z) + g(z)] = \log f(z) + \log \left[ 1 + \frac{g(z)}{f(z)} \right],$$

whence

$$\arg [f(z) + g(z)] = \arg f(z) + \arg \left[ 1 + \frac{g(z)}{f(z)} \right]. \quad (9.57)$$

On the contour  $C$ , we have

$$\left| \frac{g(z)}{f(z)} \right| < 1.$$

It thus follows that the points

$$w = 1 + \frac{g(z)}{f(z)}, \quad z \in C \quad (9.58)$$

are all situated in the interior of the circle  $|1 - w| < 1$ . Since this circle does not contain the origin, the curve (9.58) cannot surround that point. As a result, the total variation of the argument of (9.58) along  $C$  is zero. Hence, by (9.57), we have

$$\Delta_c [f(z) + g(z)] = \Delta_c [f(z)].$$

Since neither  $f(z)$  nor  $f(z) + g(z)$  has poles in  $D$ , it follows from (9.54) that these two functions have the same number of zeros in  $D$ . ♣

The application of Rouché's theorem is illustrated by the following short proof of the maximum principle. If  $f(z)$  is analytic in  $D + C$  and there is a point  $z_0$  in  $D$  such that

$$|f(z)| < |f(z_0)| \text{ for } z \in C,$$

then it follows from Rouché's theorem that the function  $f(z_0) - f(z)$  and  $f(z_0)$  have the same number of zeros in  $D$  and the function  $f(z) - f(z_0)$  has at least one zero there, namely, at  $z = z_0$ . The assumption that  $|f(z)| < |f(z_0)|$  for  $z \in C$  thus leads to a contradiction.

## Exercises

1. Let  $z_j$  be the zeros of a function  $f(z)$  that is analytic in a circular domain  $D$  and let  $f(z) \not\equiv 0$ . Each zero is counted as many times as its multiplicity. Prove that for every closed curve  $C$  in  $D$  that does not pass through a zero, the sum of winding numbers yields

$$\sum_j n(C, z_j) = \frac{1}{2\pi i} \oint_C \frac{f'(z)}{f(z)} dz. \quad (9.59)$$

**Solution:** From hypothesis, we can write  $f(z) = (z - z_1)(z - z_2) \cdots (z - z_n)g(z)$ , where  $g(z)$  is analytic and  $g(z) \neq 0$  in  $D$ . Forming the logarithmic derivative, we obtain

$$\frac{f'(z)}{f(z)} = \frac{1}{z - z_1} + \frac{1}{z - z_2} + \cdots + \frac{1}{z - z_n} + \frac{g'(z)}{g(z)}$$

for  $z \neq z_j$ , and particularly on  $C$ . Since  $g(z) \neq 0$  in  $D$ , Cauchy's theorem yields  $\oint_C g'(z)/g(z) dz = 0$ . Recalling the definition of  $n(C, z_j)$ , we set the desired result (9.59). ♣

2. Show that an analytic function in a domain  $D$  that takes only real values on the boundary  $C$  of  $D$  reduces to a constant.

**Solution:** Let  $\xi = a + ib$ ,  $b \neq 0$ , be a nonreal complex number and consider the values of  $f(z) - \xi$  for  $z \in C$ . If  $b > 0$ , say, we have  $\text{Im}[f(z) - \xi] = b > 0$  since  $f(z)$  is real on  $C$ . The values of  $f(z) - \xi$  are thus confined to the upper half-plane so that the curve described by  $f(z) - \xi$  cannot surround the origin. Hence, we have  $\Delta_C[f(z) - \xi] = 0$ . Furthermore, since  $f(z) - \xi$  is analytic in  $D + C$ , it follows from the argument principle that  $f(z) - \xi \neq 0$ , i.e.,  $f(z) \neq \xi$  in  $D$ . The same reasoning also applies to values  $\xi$  for which  $b < 0$ . We thus conclude that  $f(z)$  does not take nonreal values in  $D$ .

Next we show that the above result means that  $f(z)$  reduces to a constant. Since  $f(z)$  is analytic in  $D$ , we have

$$f'(z) = \lim_{h \rightarrow 0} \frac{f(z+h) - f(z)}{h} = \lim_{h \rightarrow 0} \frac{f(z+ih) - f(z)}{ih},$$

where  $h \rightarrow 0$  through positive values. Since  $f(z)$  is real throughout  $D$ , the first limit is real and the second limit is imaginary. They can therefore be equal only if they are both zero. Since  $z$  is arbitrary, it follows that  $f'(z) = 0$  throughout  $D$ ; hence,  $f(z) = \text{const.}$  ♣

**3.** Show that all zeros of polynomials

$$p(z) = z^n + a_{n-1}z^{n-1} + \cdots + a_1z + a_0$$

are located within the region  $|z| \leq R_0$ , where

$$R_0 = \max \{1 + |a_{n-1}|, 1 + |a_{n-2}|, \cdots, 1 + |a_1|, |a_0|\}.$$

**Solution:** Let  $f(z) = z^n$ ,  $g(z) = a_{n-1}z^{n-1} + \cdots + a_1z + a_0$ , and let  $R_k = R_0 + (1/k)$  for an arbitrary fixed  $k \in \mathbf{N}$ . Observe that

$$|a_j| \leq R_0 - 1 < R_k - 1 \quad \text{for } j = 1, 2, \cdots, n-1$$

and  $|a_0| \leq R_0 < R_k$ . Then, if  $|z| = R_k$ , we have

$$\begin{aligned} |g(z)| &\leq |a_{n-1}||z|^{n-1} + \cdots + |a_1||z| + |a_0| \\ &\leq (R_k - 1)R_k^{n-1} + \cdots + (R_k - 1)R_k + R_k = R_k^n = |f(z)|. \end{aligned}$$

In view of Rouché's theorem,  $f(z)$  and  $f(z) + g(z) = p(z)$  have the same number of zeros within the region  $|z| < R_k$ . Since  $f(z)$  has  $n$  zeros and  $p(z)$  is an  $n$ th-order polynomial, we conclude that all the zeros of  $p(z)$  have to be located within the region  $|z| < R_k$ . Finally, we take the limit  $k \rightarrow \infty$  (since  $k$  is arbitrary) to find that all the zeros of  $p(z)$  have to be located within  $|z| < R_0$ . ♣

**4.** Show that the equation  $z^3 + 3z + 1 = 0$  has solutions whose absolute values are less than 2.

**Solution:** Let  $z$  be on the circle  $|z| = 2$ . Then we have

$$|z^3| = 8 > 3 \cdot 2 + 1 > 3|z| + 1 \geq |3z + 1|.$$

This means that there are three solutions to the equation  $z^3 + 3z + 1 = 0$  and that all of them satisfy  $|z| < 2$ . ♣

## 9.5 Dispersion Relations

### 9.5.1 Principal Value Integrals

The previous sections treated contour integrals whose integrand has no pole on the contour  $C$ . If a pole is located on  $C$ , the integrand diverges at the pole so that we cannot use ordinary integration methods. This difficulty is overcome by introducing a new concept called the **principal value integral**. To derive it, we consider an integral

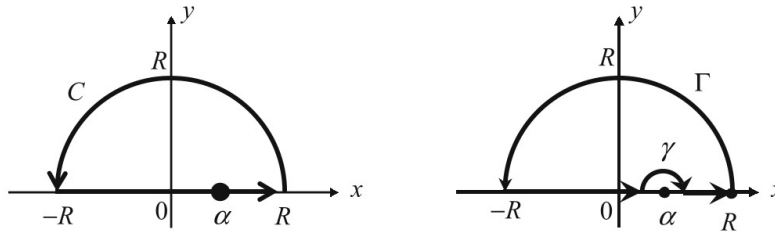
$$I = \oint_C \frac{f(z)}{z - \alpha} dz \quad (9.60)$$

with the integration contour depicted in Fig. 9.6. In (9.60),  $\alpha$  is assumed to be real without loss of generality. In addition, we assume that  $f(z)$  is analytic at  $\text{Im}z > 0$ , and behaves as  $z^\beta |f(z)| \rightarrow A$  ( $\beta > 0$ ) as  $|z| \rightarrow \infty$  there. In order for the integral (9.60) to be defined, the contour  $C$  must be traversed in such a way as to avoid the pole at  $z = \alpha$ . Then, since both  $f(z)$  and  $1/(z - \alpha)$  are analytic within and on  $C$ , (9.60) equals zero. Therefore, by breaking it up, we obtain the following expression:

$$\begin{aligned} & \oint_C \frac{f(z)}{z - \alpha} dz \\ &= \int_{-R}^{\alpha-r} \frac{f(x)}{x - \alpha} dx + \int_{\gamma} \frac{f(z)}{z - \alpha} dz + \int_{\alpha+r}^R \frac{f(x)}{x - \alpha} dx + \int_{\Gamma} \frac{f(z)}{z - \alpha} dz \\ &= 0. \end{aligned} \quad (9.61)$$

Here  $r$  is the radius of the small semicircle  $\gamma$  centered at  $x = \alpha$  and  $R$  is the radius of the large semicircle  $\Gamma$  centered at the origin. The radius  $r$  can be chosen as small as we please and  $R$  can be chosen as large as we please.

Our current interest is to determine where the sum of the four integrals appearing in the second line of (9.61) converges in the limits of  $r \rightarrow 0$  and  $R \rightarrow \infty$ . This is seen by evaluating the integrals along  $\gamma$  and  $\Gamma$  given in (9.61). First, once we set  $z = Re^{i\theta}$ , the integral along the large semicircle  $\Gamma$  yields



**Fig. 9.6.** Integration contour on which the pole of the integrand is located

$$\int_{\Gamma} \frac{f(z)}{z - \alpha} dz = i \int_0^{\pi} \frac{f(Re^{i\theta})}{Re^{i\theta} - \alpha} Re^{i\theta} d\theta;$$

hence,

$$\left| \int_{\Gamma} \frac{f(z)}{z - \alpha} dz \right| \leq \frac{R}{|R - \alpha|R^{\beta}} \int_0^{\pi} |f(Re^{i\theta})| d\theta, \quad (9.62)$$

where we have used the inequality

$$|Re^{i\theta} - \alpha| = \sqrt{R^2 + \alpha^2 - 2R\alpha \cos \theta} \geq \sqrt{R^2 + \alpha^2 - 2R\alpha} = |R - \alpha|.$$

In the limit  $R \rightarrow \infty$ , the right-hand side of (9.62) vanishes since  $\beta > 0$ . Therefore, the integral over the semicircle  $\Gamma$  can be made arbitrarily small by choosing  $R$  sufficiently large.

Next, we write the integral along  $\gamma$  as

$$\int_{\gamma} \frac{f(z)}{z - \alpha} dz = f(\alpha) \int_{\gamma} \frac{1}{z - \alpha} dz + \int_{\gamma} \frac{f(z) - f(\alpha)}{z - \alpha} dz. \quad (9.63)$$

By setting  $z - \alpha = re^{i\theta}$ , the first integral on the right-hand side is evaluated as

$$f(\alpha) \int_{\gamma} \frac{1}{z - \alpha} dz = if(\alpha) \int_{\pi}^0 d\theta = -i\pi f(\alpha).$$

In addition, the Taylor series expansion of  $f(z)$  around  $z = \alpha$  yields

$$\frac{f(z) - f(\alpha)}{z - \alpha} dz = f'(\alpha) \cdot i\varepsilon e^{i\theta} d\theta + \frac{f''(\alpha)}{2} \cdot \varepsilon e^{i\theta} \cdot i\varepsilon e^{i\theta} d\theta + \cdots = O(\varepsilon),$$

which means that the second integral in (9.63) vanishes in the limit  $r \rightarrow 0$ . Equation (9.61) thus yields

$$\lim_{R \rightarrow \infty} \lim_{r \rightarrow 0} \left[ \int_{-R}^{\alpha-r} \frac{f(x)}{x - \alpha} dx + \int_{\alpha+r}^R \frac{f(x)}{x - \alpha} dx \right] - i\pi f(\alpha) = 0. \quad (9.64)$$

Now we introduce a new notation as shown below.

#### ♠ Principal value integral:

The notation

$$\mathcal{P} \int_{-R}^R \frac{f(x)}{x - \alpha} dx \equiv \lim_{r \rightarrow 0} \left[ \int_{-R}^{\alpha-r} \frac{f(x)}{x - \alpha} dx + \int_{\alpha+r}^R \frac{f(x)}{x - \alpha} dx \right]$$

provides the **principal value integral** (or the **Cauchy principal value**) of  $f(z)/(z - \alpha)$  for real  $\alpha$ .

with this notation, (9.64) reads

$$\lim_{R \rightarrow \infty} \mathcal{P} \int_{-R}^R \frac{f(x)}{x - \alpha} dx = i\pi f(\alpha),$$

where  $f(x)$  is a complex-valued function of a real variable  $x$ . For the sake of brevity, we write this simply as

$$\mathcal{P} \int_{-\infty}^{\infty} \frac{f(x)}{x - \alpha} dx = i\pi f(\alpha). \quad (9.65)$$

This result provides a way to evaluate the contour integrals involving singularities on the integration path. When we decompose  $f(x)$  in (9.65) as  $f(x) = f_R(x) + if_I(x)$  and equate the real and imaginary parts, we obtain an important relation between  $f_R$  and  $f_I$ :

♠ **Hilbert transform pair:**

A pair of functions  $f_R$  and  $f_I$  that satisfies the relations

$$\begin{aligned} f_R(\alpha) &= \frac{1}{\pi} \mathcal{P} \int_{-\infty}^{\infty} \frac{f_I(x)}{x - \alpha} dx, \\ f_I(\alpha) &= -\frac{1}{\pi} \mathcal{P} \int_{-\infty}^{\infty} \frac{f_R(x)}{x - \alpha} dx. \end{aligned} \quad (9.66)$$

is called a **Hilbert transform pair**.

It readily follows from (9.66) that if  $f_I(x) \equiv 0$ , then  $f_R(x) \equiv 0$ .

### 9.5.2 Several Remarks

The principal value integral is seen as a way of avoiding singularities on a path of integration; we integrate to the point just before the singularity in question, skip over the singularity, and begin integrating again immediately beyond the singularity. This prescription enables us to make sense out of integrals such as

$$\int_{-R}^R \frac{dx}{x}. \quad (9.67)$$

Apparently, this integral seems to be zero, since an odd function is integrated over a symmetric domain. However, the singularity at the origin makes the integral meaningless unless we insert a symbol  $\mathcal{P}$  in front of it. Following the prescription for principal value integrals, we can easily evaluate the principal value of (9.67):

$$\mathcal{P} \int_{-R}^R \frac{dx}{x} = \lim_{r \rightarrow 0} \left[ \int_{-R}^{-r} \frac{dx}{x} + \int_r^R \frac{dx}{x} \right].$$

In the first integral on the right-hand side, we set  $x = -y$ . Then

$$\mathcal{P} \int_{-R}^R \frac{dx}{x} = \lim_{r \rightarrow 0} \left[ \int_R^r \frac{dy}{y} + \int_r^R \frac{dx}{x} \right],$$

where the two integrals within the brackets obviously cancel out. Consequently, we have

$$\mathcal{P} \int_{-R}^R \frac{dx}{x} = 0. \quad (9.68)$$

We emphasize again that the integral (9.68) is completely different from the meaningless quantity in (9.67).

As a further step, we evaluate the principal value integral defined by

$$\mathcal{P} \int_{-R}^R \frac{f(x)}{x - \alpha} dx.$$

It follows from the result of (9.83) that

$$\begin{aligned} \mathcal{P} \int_{-R}^R \frac{f(x)}{x - \alpha} dx &= \mathcal{P} \int_{-R}^R \left[ \frac{f(\alpha)}{x - \alpha} + \frac{f(x) - f(\alpha)}{x - \alpha} \right] dx \\ &= f(\alpha) \ln \left( \frac{R - \alpha}{R + \alpha} \right) + \mathcal{P} \int_{-R}^R \frac{f(x) - f(\alpha)}{x - \alpha} dx. \end{aligned} \quad (9.69)$$

It often happens that the second integral in the second equation in (9.69) is not be singular at  $x = \alpha$ ; for instance, as in the case where  $f(x)$  is differentiable at  $x = \alpha$ . In this case, the symbol  $\mathcal{P}$  there can be dropped

Particularly interesting is the behavior of (9.69) in the limit  $R \rightarrow \infty$ , which yields

$$\mathcal{P} \int_{-\infty}^{\infty} \frac{f(x)}{x - \alpha} dx = \mathcal{P} \int_{-\infty}^{\infty} \frac{f(x) - f(\alpha)}{x - \alpha} dx. \quad (9.70)$$

Hence, substituting (9.70) into (9.65), we obtain

$$\begin{aligned} f_R(\alpha) &= \frac{1}{\pi} \mathcal{P} \int_{-\infty}^{\infty} \frac{f_I(x) - f_I(\alpha)}{x - \alpha} dx, \\ f_I(\alpha) &= -\frac{1}{\pi} \mathcal{P} \int_{-\infty}^{\infty} \frac{f_R(x) - f_R(\alpha)}{x - \alpha} dx, \end{aligned} \quad (9.71)$$

which are complementary expressions of a Hilbert transform pair (9.66).

*Remark.* Equation (9.70) is equivalent to

$$\mathcal{P} \int_{-\infty}^{\infty} \frac{f(a)}{x - a} dx = 0, \quad \text{and thus} \quad \mathcal{P} \int_{-\infty}^{\infty} \frac{dx}{x - a} = 0,$$

which readily follows from the result (9.68).

### 9.5.3 Dispersion relations

Mathematical arguments given so far are interesting in their own right, but their applications to physical sciences are also significant. In the following discussions, we show that general physical quantities associated with response phenomena satisfy the Hilbert transform relations given in (9.66) and (9.71). In the language of physics, the relation between corresponding parts of Hilbert transform pairs referred to as a **dispersion relation**, plays an important role in describing the properties of response functions.

We begin by considering a physical system for which an input  $I(t)$  is related to a response  $R(t)$  in the following linear manner:

$$R(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} G(t-t')I(t')dt'. \quad (9.72)$$

For example,  $I(t')$  might be the electric field acting on a physical object at a time  $t'$  and  $R(t)$  is the resulting polarization field at time  $t$ . We have assumed that  $G$  depends only on the difference  $t-t'$  because we want the system to respond to a sharp input at  $t_0$  as expressed by  $I(t') = I_0\delta(t'-t_0)$ . In the same way, it would respond to a sharp input at  $t_0 + \tau$ , i.e., at a time  $\tau$  later. For the first case, we have

$$R_1(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} G(t-t')I_0\delta(t'-t_0)dt' = \frac{I_0}{\sqrt{2\pi}}G(t-t_0). \quad (9.73)$$

and for the second,

$$R_2(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} G(t-t')I_0\delta(t'-t_0-\tau)dt' = \frac{I_0}{\sqrt{2\pi}}G(t-t_0-\tau),$$

or, in other words,

$$R_2(t+\tau) = \frac{I_0}{\sqrt{2\pi}}G(t-t_0) = R_1(t).$$

Thus if we shift the input by  $\tau$ , the response is also shifted by  $\tau$ .

Now, in order to derive the dispersion relation for the physical systems of interest, we consider the Fourier transform of (9.72). Using the convolution theorem, we find that

$$r(\omega) = g(\omega)j(\omega),$$

where

$$r(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} R(t)e^{i\omega t}dt, \quad g(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} G(t)e^{i\omega t}dt,$$

$$\text{and} \quad j(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} I(t)e^{i\omega t}dt.$$

Notably, it is possible to extend  $g(\omega)$  into the complex  $z$ -plane, based on the assumptions that

- (i)  $g(z)$  is analytic for  $\text{Im}z > 0$ , and
- (ii)  $g(z) \rightarrow 0$  as  $z \rightarrow \infty$ .

Observe that (i) and (ii), are the conditions under which we derived the Hilbert transform pair (see Sect. 9.4.1). After some discussion, we see that  $g(z)$  arising from a  $G(t)$  that satisfies the necessary assumptions yields

$$\begin{aligned} g_R(\omega) &= \frac{1}{\pi} \mathcal{P} \int_{-\infty}^{\infty} \frac{g_I(\omega')}{\omega' - \omega} d\omega', \\ g_I(\omega) &= -\frac{1}{\pi} \mathcal{P} \int_{-\infty}^{\infty} \frac{g_R(\omega')}{\omega' - \omega} d\omega'. \end{aligned} \quad (9.74)$$

These relations between  $g_R$  and  $g_I$  are called the **dispersion relations** for  $g$ . The validity of assumptions (i) and (ii) that the function  $g(z)$  must satisfy is demonstrated in Sect. 9.5.6.

#### 9.5.4 Kramers–Kronig Relations

The term “dispersion relation” is often restricted to mean a relation between two functions whose arguments are quantitatively treatable experimentally. For instance, in (9.74) only a positive frequency ( $\omega \geq 0$ ) should actually be accessible, so they are not directly practical as they stand. In the following, we derive an alternative expression of the dispersion relations that involve only positive, experimentally meaningful frequencies.

We first assume that  $G(t)$  is real, which is obvious from (9.73), where  $R_1$  and  $I_0$  are real. Hence, we may proceed as follows:

$$\begin{aligned} g(z) &= \frac{1}{\sqrt{2\pi}} \int_0^{\infty} G(t) e^{izt} dt, \\ g^*(z) &= \frac{1}{\sqrt{2\pi}} \int_0^{\infty} G^*(t) e^{-iz^*t} dt = \frac{1}{\sqrt{2\pi}} \int_0^{\infty} G(t) e^{-iz^*t} dt \\ &= g(-z^*). \end{aligned} \quad (9.75)$$

As a consequence, we have

$$g^*(z) = g(-z^*),$$

which is referred to as the **reality condition**.

Next let us assume  $z$  to be real ( $z = \omega$ ) in order to discuss the behavior of  $g(z)$  on the real axis. It follows from the reality condition (9.75) that

$$g_R(\omega) - ig_I(\omega) = g_R(-\omega) + ig_I(-\omega)$$

or

$$g_R(\omega) = g_R(-\omega) \quad \text{and} \quad g_I(\omega) = -g_I(-\omega). \quad (9.76)$$

That is,  $g_R$  and  $g_I$  are even and odd functions of  $\omega$ , respectively. Note that if the conditions in (9.76) are satisfied, the function

$$G(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(\omega) e^{-i\omega t} d\omega$$

becomes a real function. (The proof is left to the reader).

Now we rewrite the first part of (9.74) as

$$g_R(\omega) = \frac{1}{\pi} \mathcal{P} \int_{-\infty}^0 \frac{g_I(\omega')}{\omega' - \omega} d\omega' + \frac{1}{\pi} \mathcal{P} \int_0^{\infty} \frac{g_I(\omega')}{\omega' - \omega} d\omega'.$$

we rewrite  $\omega' \rightarrow -\omega'$  in the first integral and use (9.76) to obtain

$$g_R(\omega) = \frac{2}{\pi} \mathcal{P} \int_0^{\infty} \frac{\omega' g_I(\omega')}{\omega'^2 - \omega^2} d\omega' \quad (9.77)$$

and an identical procedure yields

$$g_I(\omega) = -\frac{2\omega}{\pi} \mathcal{P} \int_0^{\infty} \frac{g_R(\omega')}{\omega'^2 - \omega^2} d\omega'. \quad (9.78)$$

Eventually, the expressions (9.77) and (9.78) involve only positive, experimentally accessible frequencies. These equations are referred to as the **Kramers–Kronig relations**.

### 9.5.5 Subtracted Dispersion Relation

In deriving dispersion relations, it often happens that the quantity of interest, say  $g(z)$ , does not tend toward zero as  $|z| \rightarrow \infty$ . Furthermore, we are not usually fortunate enough to know the precise behavior of the quantity as  $|z|$  tends to infinity. Nevertheless, if we at least know that the quantity is bounded for large values of  $|z|$ , the dispersion relation can be reformulated in the following way:

Suppose that  $f(z)$  is analytic in the upper half-plane, and let  $\alpha_0$  be some point on the real axis at which  $f(z)$  is analytic. Our aim is to derive the dispersion relation for  $f(x)$  under the condition that the asymptotic behavior of  $f(z)$  for  $z \rightarrow \infty$  is unknown. Then, instead of  $f(z)$ , we consider the function

$$\frac{f(z) - f(\alpha_0)}{z - \alpha_0} \equiv \phi(z),$$

which is also analytic in the upper half-plane and not singular at  $z = \alpha_0$ , and  $|\phi(z)| \rightarrow 0$  as  $|z| \rightarrow \infty$  owing to the boundedness of  $|f(z)|$  for  $z \rightarrow \infty$ . Thus in a manner, similar to the case in (9.65), we can write

$$i\pi\phi(x) = \mathcal{P} \int_{-\infty}^{\infty} \frac{\phi(x')}{x' - x} dx'.$$

In actuality, we have

$$\begin{aligned}
& i\pi \left[ \frac{f(x) - f(\alpha_0)}{x - \alpha_0} \right] \\
&= \mathcal{P} \int_{-\infty}^{\infty} \frac{f(x') - f(\alpha_0)}{(x' - x)(x' - \alpha_0)} dx' \\
&= \mathcal{P} \int_{-\infty}^{\infty} \frac{f(x')}{(x' - x)(x' - \alpha_0)} dx' - \frac{f(\alpha_0)}{x - \alpha_0} \mathcal{P} \int_{-\infty}^{\infty} \left( \frac{1}{x' - x} - \frac{1}{x' - \alpha_0} \right) dx',
\end{aligned}$$

so

$$\begin{aligned}
i\pi f(x) = & i\pi f(\alpha_0) + (x - \alpha_0) \mathcal{P} \int_{-\infty}^{\infty} \frac{f(x')}{(x' - x)(x' - \alpha_0)} dx' \\
& - f(\alpha_0) \mathcal{P} \int_{-\infty}^{\infty} \frac{dx'}{x' - x} + f(\alpha_0) \mathcal{P} \int_{-\infty}^{\infty} \frac{dx'}{x' - \alpha_0}.
\end{aligned}$$

The last two principal value integrals are equal to zero as we demonstrate later in (9.83). Hence, separating the real and imaginary parts, we finally obtain

$$\begin{aligned}
f_R(x) &= f_R(\alpha_0) + \frac{x - \alpha_0}{\pi} \mathcal{P} \int_{-\infty}^{\infty} \frac{f_I(x')}{(x' - x)(x' - \alpha_0)} dx', \\
f_I(x) &= f_I(\alpha_0) - \frac{x - \alpha_0}{\pi} \mathcal{P} \int_{-\infty}^{\infty} \frac{f_R(x')}{(x' - x)(x' - \alpha_0)} dx'. \quad (9.79)
\end{aligned}$$

Relations of the type of (9.79) are referred to as **once-subtracted dispersion relations**. Emphasis is placed on the fact that the relations (9.79) are free from the assumption that  $|f(z)|$  should vanish in the limit  $z \rightarrow \infty$ . For them to be of use in a particular physical problem, we must have a means of determining, say,  $f_R(\alpha_0)$  for some  $\alpha_0$ .

### 9.5.6 Derivation of Dispersion Relations

This subsection provides a proof of the dispersion relation (9.74). We shall see that by making a few very reasonable assumptions about the system in question, we can show that the real and imaginary parts of the physical quantity  $g(\omega)$  are intimately related to one another for real values of  $\omega$  (i.e., a dispersion relation). The key assumption is the **causality requirement**: we may say that causality of the function  $G(t)$  implies the analytic properties of  $g(z)$  in the upper half-plane and thus verifies the dispersion relations with respect to  $g(\omega)$  on the real axis.

Toward this end, let us consider what can be said about  $G(\tau)$  on general physical grounds. First to be noted is that an input at  $t$  should not give rise to a response at times prior to  $t$ , i.e.,  $G(\tau) = 0$  for  $\tau < 0$ . Thus we have

$$R(t) = \int_{-\infty}^t G(t - t') I(t') dt', \quad (9.80)$$

which shows that the response at  $t$  is the weighted linear superposition of all inputs prior to  $t$ , which is the causality requirement.

Secondly, the possibility that  $G(\tau)$  is singular for any finite  $\tau$  is excluded because, on physical grounds, the response from a sharp input given by

$$R(t) = \frac{I_0}{\sqrt{2\pi}} G(t - t_0), \quad t > t_0$$

must always be finite.

Finally, it is assumed that the effect of an input in the remote past does not appreciably influence the present. This may be stated as the requirement that  $G(\tau) \rightarrow 0$  as  $\tau \rightarrow \infty$ , since the response to any impulse dies down after a sufficiently long time (i.e., any system has some dissipative mechanism). Furthermore,  $G(\tau)$  should vanish faster than  $\tau^{-1}$  so that it becomes integrable. Recall that  $g(z)$  is defined through an integration of  $G(t)$  with respect to  $t$ .

The following three points summarise our physically motivated assumptions on  $G(\tau)$ :

- (i)  $G(\tau) = 0$  for  $\tau < 0$ ,
- (ii)  $G(\tau)$  is bounded for all  $\tau$ , and
- (iii)  $|G(\tau)|$  is integrable, so  $G(\tau) \rightarrow 0$  faster than  $1/\tau$  as  $\tau \rightarrow \infty$ .

We demonstrate below that these three assumptions for  $G(t)$  lead naturally to the two conditions for  $g(z)$  under which we have derived the dispersion relation of  $g(\omega)$ .

First, we show that these three conditions require that  $|g(z)| \rightarrow 0$  at  $z \rightarrow \infty$  on the upper half-plane. It is possible to write

$$g(\omega) = \frac{1}{\sqrt{2\pi}} \int_0^\infty G(t) e^{i\omega t} dt.$$

We extend this relation into the complex plane by using the definition

$$g(z) = \frac{1}{\sqrt{2\pi}} \int_0^\infty G(t) e^{izt} dt = \frac{1}{\sqrt{2\pi}} \int_0^\infty G(t) e^{i\omega t} e^{-\eta t} dt,$$

where we have written  $z = \omega + i\eta$ . We now restrict our attention to the upper half-plane ( $\eta > 0$ ), where the term  $e^{-\eta t}$  is a decaying exponential. For  $0 < \theta < \pi$ , it reads

$$|g(z)| \leq \frac{1}{\sqrt{2\pi}} M \int_0^\infty e^{-(|z| \sin \theta)t} dt,$$

where we have replaced  $G(t)$  by its maximum value  $M$  in view of assumption (ii) above. Hence, we have

$$|g(z)| \leq \frac{M_G}{\sqrt{2\pi}|z| \sin \theta}.$$

This means that for  $0 < \theta < \pi$ ,  $|g(z)| \rightarrow 0$  as  $|z| \rightarrow \infty$ . On the other hand, when  $\theta = 0$  or  $\pi$ , we have

$$g(\omega, \eta = 0) = \frac{1}{\sqrt{2\pi}} \int_0^\infty G(t) e^{i\omega t} dt.$$

This results in Parseval's identity:

$$\int_{-\infty}^\infty |g(\omega, \eta = 0)|^2 d\omega = \frac{1}{2\pi} \int_{-\infty}^\infty |G(t)|^2 dt,$$

where both sides of improper integrals converge. (See Sect. 3.4.2 for the convergence conditions of an improper integral.) Thus  $|g(\omega, \eta = 0)|$  vanishes as  $\omega \rightarrow \infty$ . As a result,  $|g(z)| \rightarrow 0$  as  $|z| \rightarrow \infty$  in the whole region of  $0 \leq \theta \leq \pi$ , i.e., in any direction in the upper half-plane.

Now we want to show that  $g(z)$  is analytic in the upper half-plane. Using

$$g(z) = \frac{1}{\sqrt{2\pi}} \int_0^\infty G(t) e^{izt} dt = \frac{1}{\sqrt{2\pi}} \int_0^\infty G(t) e^{i\omega t} e^{-\eta t} dt, \quad (9.81)$$

we see that for  $\eta > 0$ ,

$$\frac{d^n g}{dz^n} = \frac{1}{\sqrt{2\pi}} \int_0^\infty G(t) \frac{d^n}{dz^n} e^{izt} dt = \frac{i^n}{\sqrt{2\pi}} \int_0^\infty t^n G(t) e^{i\omega t} e^{-\eta t} dt. \quad (9.82)$$

The integrals in (9.82) are uniformly convergent owing to the term  $e^{-\eta t}$  ( $\eta > 0$ ,  $t > 0$ ). Thus  $g(z)$  is analytic in the upper half-plane ( $\eta > 0$ ). Hence, for any  $g(z)$  arising from a  $G(t)$  that satisfies assumptions (i), (ii), and (iii), we can proceed according to the argument in Sect. 9.4.1, and we finally obtain the dispersion relation (9.74).

## Exercises

1. Prove that

$$\mathcal{P} \int_{-R}^R \frac{dx}{x-a} = \ln \left( \frac{R-a}{R+a} \right)$$

when  $-R < a < R$ .

**Solution:** We write

$$\mathcal{P} \int_{-R}^R \frac{dx}{x-a} = \lim_{\varepsilon \rightarrow 0} \left[ \int_{-R}^{a-\varepsilon} \frac{dx}{x-a} + \int_{\varepsilon+a}^R \frac{dx}{x-a} \right].$$

Setting  $x = -y$  in the first integral on the right-hand side, we find that

$$\begin{aligned} \mathcal{P} \int_{-R}^R \frac{dx}{x-a} &= \lim_{\varepsilon \rightarrow 0} \left[ \int_R^{\varepsilon-a} \frac{dy}{y+a} + \ln(R-a) - \ln \varepsilon \right] \\ &= \lim_{\varepsilon \rightarrow 0} [\ln \varepsilon - \ln(R+a) + \ln(R-a) - \ln \varepsilon] \\ &= \ln \left( \frac{R-a}{R+a} \right) \quad (-R < a < R). \quad \clubsuit \quad (9.83) \end{aligned}$$

2. By using the formula (9.71), prove that

$$\int_{-\infty}^{\infty} \frac{\sin x}{x} dx = \pi.$$

**Solution:** Consider the function  $f(z) = e^{iz}$ . This function is analytic everywhere, and if we write  $z = Re^{i\theta}$ , then  $|f(z)| \rightarrow 0$  as  $R \rightarrow \infty$  for all  $\theta$  such that  $0 < \theta < \pi$ . In this case,  $f_R(x) = \cos x$  and  $f_I(x) = \sin x$ , so using (9.71), we obtain

$$\cos \alpha = (1/\pi) \int_{-\infty}^{\infty} (\sin x - \sin \alpha)/(x - \alpha) dx.$$

Since  $\sin x - \sin \alpha = 2 \sin[(x-\alpha)/2] \cos[(x+\alpha)/2]$ , there is no singularity of the integrand at  $x = \alpha$ . For the special case  $\alpha = 0$ , we find that  $1 = (1/\pi) \int_{-\infty}^{\infty} (\sin x/x) dx$ , i.e.,  $\int_{-\infty}^{\infty} (\sin x/x) dx = \pi$ . From this result, we also obtain  $\int_0^{\infty} (\sin x/x) dx = \pi/2$  by symmetry. ♣

3. Show that the integral  $S(t) = \frac{1}{2\pi i} \lim_{\varepsilon \rightarrow 0} \int_{-\infty}^{\infty} \frac{e^{ixt}}{x - i\varepsilon} dx$  reads

$$S(t) = \begin{cases} 1, & t > 0, \\ 0, & t < 0. \end{cases}$$

**Solution:** Taking the contours  $\text{Im}(z) > 0$  for  $t > 0$ , and  $\text{Im}(z) < 0$  for  $t < 0$ , we have the desired result, which is the integral representation of Heaviside's step function. ♣



## Conformal Mapping

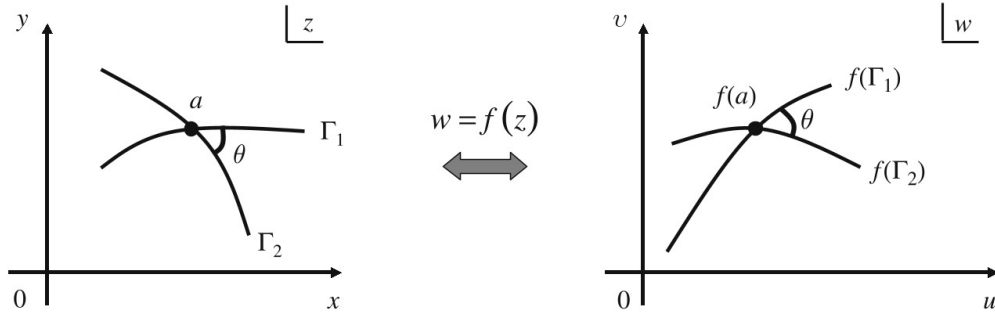
**Abstract** Conformal mapping refers to transformation from one complex plane to another such that the local angles and shapes of infinitesimally small figures are preserved. This special class of mapping is indispensable for solving physics and engineering problems that are expressed in terms of complex functions with inconvenient geometries. In this chapter we show that a problem can be drastically simplified by choosing an appropriate mapping, which allows us to evaluate the solution using elementary calculus.

### 10.1 Fundamentals

#### 10.1.1 Conformal Property of Analytic Functions

We are concerned here the mapping properties of an **analytic function**  $w = f(z)$  in a domain  $D$  on the  $z$ -plane into the  $w$ -plane. Through the mapping, any line drawn on the  $z$ -plane results in a line on the  $w$ -plane. Particularly when  $f = u + iv$  is analytic, the transformation is **angle-preserving** or **conformal**. This means that through the transformation from  $(x, y)$  to  $(u, v)$ , the angle between the crossing lines on the  $w$ -plane is equal to the angle between the crossing lines on the  $z$ -plane (see Fig. 10.1). In physics and engineering, the subject derives its usefulness from the possibility of transforming a problem that occurs naturally in a rather difficult setting into another simpler one.

Let  $D$  be a domain on the  $z$ -plane, and let  $\Gamma_1$  and  $\Gamma_2$  be two differentiable arcs lying in  $D$  and intersecting at a point  $z = a$  in  $D$ . If  $f(z)$  is an analytic function in  $D$ , the images  $f(\Gamma_1)$  and  $f(\Gamma_2)$  are differentiable arcs lying in a domain  $D' = f(D)$  and intersecting at a point  $a' = f(a)$ . Then we say the following:



**Fig. 10.1.** Angle-preserving property of a conformal mapping  $w = f(z)$

♠ **Conformal mapping:**

The mapping  $w = f(z)$  is **conformal** at  $z = a$  if for every such pair of arcs, the angle between the arcs  $\Gamma_1$  and  $\Gamma_2$  intersecting at  $z = a$  on the  $z$ -plane is equal to the angle between the arcs  $f(\Gamma_1)$  and  $f(\Gamma_2)$  at their intersecting point  $f(a)$  on the  $w$ -plane.

The mapping is said to be conformal in  $D$  if it is conformal at each point in  $D$ . We shall see that if a function  $w = f(z)$  is analytic, it is necessarily conformal except at a finite number of specific points; this fact is formally stated below.

♠ **Theorem:**

Given an analytic function  $f(z)$ , the mapping  $w = f(z)$  is conformal at  $z = a$  if and only if  $f'(a) \neq 0$ .

*Proof* For proving sufficiency, we consider the arcs  $\Gamma_1$  and  $\Gamma_2$  given parametrically by

$$z_1 = \Psi_1(t) \quad \text{and} \quad z_2 = \Psi_2(t) \quad (0 \leq t \leq 1)$$

and assume that  $z_1, z_2$  are points on  $\Gamma_1, \Gamma_2$  at a short distance  $\ell$  from  $z = a$ . Then, from the relation

$$z_1 - a = \ell e^{i\alpha}, \quad z_2 - a = \ell e^{i\beta},$$

we have the ratio

$$\frac{z_2 - a}{z_1 - a} = e^{i(\beta - \alpha)}.$$

As  $\ell \rightarrow 0$ ,  $\beta - \alpha$  must approach the angle  $\theta$  between the curves on the  $z$ -plane. That is,

$$\theta = \lim_{\ell \rightarrow 0} \arg \left( \frac{z_2 - a}{z_1 - a} \right).$$

For the angle  $\tilde{\theta}$  between the arcs of  $f(\Gamma_1)$  and  $f(\Gamma_2)$  at  $f(a)$ , we have

$$\begin{aligned}\tilde{\theta} &= \lim_{\ell \rightarrow 0} \arg \left[ \frac{f(z_2) - f(a)}{f(z_1) - f(a)} \right] \\ &= \lim_{\ell \rightarrow 0} \arg \left[ \frac{\frac{f(z_2) - f(a)}{z_2 - a} \cdot (z_2 - a)}{\frac{f(z_1) - f(a)}{z_1 - a} \cdot (z_1 - a)} \right] \\ &= \lim_{\ell \rightarrow 0} \arg \left[ \frac{f'(a) \cdot (z_2 - a)}{f'(a) \cdot (z_1 - a)} \right] = \theta, \quad \text{if } f'(a) \neq 0. \quad (10.1)\end{aligned}$$

Thus, the condition  $f'(a) \neq 0$  is necessary. Conversely if  $f^{(n)}(a) = 0$  with  $n = 1, 2, \dots$  and  $f^{(p)}(a) \neq 0$ , near  $z = a$  we have

$$f(z) = f(a) + O[(z - a)^p].$$

Thus, we get

$$\begin{aligned}\tilde{\theta} &= \lim_{\ell \rightarrow 0} \arg \left[ \frac{f(z_2) - f(a)}{f(z_1) - f(a)} \right] \\ &= \lim_{\ell \rightarrow 0} \arg \left[ \frac{(z_2 - a)^p}{(z_1 - a)^p} \right] \\ &= p \lim_{\ell \rightarrow 0} \arg \left( \frac{z_2 - a}{z_1 - a} \right) = p\theta,\end{aligned}$$

which shows that the angle is magnified by  $p$ . Therefore, if the mapping  $w = f(z)$  is conformal, we necessarily have  $p = 1$ , which completes the proof of the sufficiency of the condition. ♣

### 10.1.2 Scale Factor

There is another important geometric property that analytic functions possess: whenever  $f(z)$  is analytic, any infinitesimal figure plotted on the  $z$ -plane is transformed into a similar figure on the  $w$ -plane with a change in size but with the proportions (and angles) preserved. We prove this by considering the length of an infinitesimally small quantity  $df$  given by

$$df = du + idv = \left( \frac{\partial u}{\partial x} dx + \frac{\partial u}{\partial y} dy \right) + i \left( \frac{\partial v}{\partial x} dx + \frac{\partial v}{\partial y} dy \right). \quad (10.2)$$

Its square length reads

$$\begin{aligned}
 |df|^2 &= \left( \frac{\partial u}{\partial x} dx + \frac{\partial u}{\partial y} dy \right)^2 + \left( \frac{\partial v}{\partial x} dx + \frac{\partial v}{\partial y} dy \right)^2 \\
 &= \left[ \left( \frac{\partial u}{\partial x} \right)^2 + \left( \frac{\partial v}{\partial x} \right)^2 \right] (dx)^2 + \left[ \left( \frac{\partial u}{\partial y} \right)^2 + \left( \frac{\partial v}{\partial y} \right)^2 \right] (dy)^2 \\
 &\quad + 2 \left( \frac{\partial u}{\partial x} \frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \frac{\partial v}{\partial y} \right) dx dy.
 \end{aligned} \tag{10.3}$$

Substituting the Cauchy–Riemann relations into (10.3), we obtain

$$|df|^2 = h^2 |dz|^2, \quad \text{where } h = \sqrt{\left( \frac{\partial u}{\partial x} \right)^2 + \left( \frac{\partial u}{\partial y} \right)^2} = \sqrt{\left( \frac{\partial v}{\partial x} \right)^2 + \left( \frac{\partial v}{\partial y} \right)^2}. \tag{10.4}$$

The quantity  $h$  is known as a **scale factor** and measures a magnification ratio of the elementary lines through the transformation  $w = f(z)$ . From (10.4), it readily follows that

$$h = \left| \frac{df}{dz} \right|. \tag{10.5}$$

We see from (10.5) that since  $df/dz$  is isotropic, the scale factor  $h$  is also isotropic (i.e., independent of the direction of  $dz$ ) for any analytic function  $f$ . This means that any infinitesimal figures on the  $z$ -plane are transformed into similar figures on the  $w$ -plane with a change in their size by  $h = |df/dz|$ .

Note that the magnitude of  $h$  depends on points  $z$  and may vanish at points where  $f'(z) = 0$ . Points where  $f'(z) = 0$  are called **critical points** of the transformation  $w = f(z)$ , and at these points, the transform becomes non conformal. The simplest example is

$$f(z) = z^2$$

for which we have

$$h = |f'(0)| = 0.$$

In fact, when two line elements passing through  $z = 0$  make an angle  $\beta - \alpha$  with respect to one another, the corresponding lines on the  $w$ -plane make an angle of  $2(\beta - \alpha)$ . Thus mapping is not conformal at  $z = 0$ . In general, the region in the neighborhood of the point at which  $h = 0$  on the  $w$ -plane becomes greatly compressed. In contrast, the corresponding region on the  $z$ -plane is tremendously expanded.

### 10.1.3 Mapping of a Differential Area

The scale factor  $h$  given in (10.4) can be derived in a different way by considering the conformal mapping of a differential area. Let  $f(z)$  be a conformal mapping that transforms any points in  $D$  of the  $z$ -plane onto a region  $S$  of the

$w$ -plane. In the domain  $D$ , we define a rectangular differential area element with sides of the rectangle parallel to the  $x$  and  $y$ -axes. These sides are given by

$$dz_1 = dx \quad \text{and} \quad dz_2 = idy,$$

The images of  $dz_1$  and  $dz_2$  are differential curves in the  $w$ -plane given by

$$dw_1 = du_1 + idv_1 \quad \text{and} \quad dw_2 = du_2 + idv_2.$$

Note that the differential area element of the rectangle in the  $z$ -plane reads  $dA_z = dxdy$  and that of the parallelogram in the  $w$ -plane is

$$dA_w = |\operatorname{Im}(dw_1^* dw_2)|.$$

Since  $dz_1 = dx$  and  $dz_2 = idy$ , the images of these line elements can be written as

$$dw_1 = \frac{\partial f}{\partial x} dx = \left( \frac{\partial u}{\partial x} + i \frac{\partial v}{\partial x} \right) dx$$

and

$$dw_2 = \frac{1}{i} \frac{\partial f}{\partial y} idy = \left( \frac{\partial u}{\partial y} + i \frac{\partial v}{\partial y} \right) dy.$$

Therefore,  $dA_w$  is given by

$$dA_w = |\operatorname{Im}(dw_1^* dw_2)| = \left| \frac{\partial u}{\partial x} \frac{\partial v}{\partial y} - \frac{\partial u}{\partial y} \frac{\partial v}{\partial x} \right| dxdy = \frac{\partial(u, v)}{\partial(x, y)} dA_z, \quad (10.6)$$

where

$$\frac{\partial(u, v)}{\partial(x, y)} \equiv \left| \frac{\partial u}{\partial x} \frac{\partial v}{\partial y} - \frac{\partial u}{\partial y} \frac{\partial v}{\partial x} \right| = \left| \begin{array}{cc} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{array} \right|$$

is called the **Jacobian determinant** of the transformation. Since  $f(z)$  is analytic,  $u$  and  $v$  satisfy the Cauchy–Riemann relations over the region  $R$ , so the Jacobian determinant can be written as

$$\frac{\partial(u, v)}{\partial(x, y)} = \left( \frac{\partial u}{\partial x} \right)^2 + \left( \frac{\partial u}{\partial y} \right)^2 = \left( \frac{\partial v}{\partial x} \right)^2 + \left( \frac{\partial v}{\partial y} \right)^2. \quad (10.7)$$

This provides a physical interpretation of the Jacobian determinant  $\partial(u, v)/\partial(x, y)$ ; namely, it is identical to the square of the same factor  $h$  introduced in (10.4).

#### 10.1.4 Mapping of a Tangent Line

We consider the mapping of a tangent line. Let  $C$  be a curve in the  $z$ -plane and  $\Gamma$  be the image of  $C$  in the  $w$ -plane (see Fig. 10.2). A differential segment  $dw$  along  $\Gamma$  is related to the differential segment  $dz$  along  $C$  by

$$dw = \frac{df}{dz} dz = f'(z) dz. \quad (10.8)$$

We suppose  $w_0$  to be a point on  $\Gamma$  that is the image of  $z_0$  on  $C$ . Then from (10.8), the tangent to  $\Gamma$  at  $w_0$ , denoted by  $\tau(w_0)$ , is related to the tangent to  $C$  at  $z_0$ , denoted by  $t(z_0)$ :

$$\tau(w_0) \equiv \left. \frac{dw}{d\lambda} \right|_{w=w_0} = f'(z_0) \left. \frac{dz}{d\lambda} \right|_{z=z_0} \equiv f'(z_0) t(z_0), \quad (10.9)$$

where  $\lambda$  parametrizes the curve of  $\Gamma$  on the  $w$ -plane.

An immediate consequence of equation (10.9) is that if  $f'(z_0) = 0$ , the tangent  $t(z_0)$  on the  $z$ -plane cannot be related to the tangent  $\tau(w_0)$  on the  $w$ -plane. The point  $z_0$  that satisfies  $f'(z) = 0$  is called a **critical point** on the curve. For simplicity in the following discussion, we assume that the curve  $C$  does not contain any critical points.

The characteristics of the mapping (10.9) become clear by employing the polar form.

$$\tau(w_0) = |\tau(w_0)| e^{i\psi(w_0)}, \quad f'(z_0) = |f'(z_0)| e^{i\phi(z_0)}, \quad \text{and} \quad t(z_0) = |t(z_0)| e^{i\theta(z_0)}. \quad (10.10)$$

The first equation shows that  $\tau(w_0)$  is oriented at an angle  $\psi(w_0)$  to the  $u$ -axis; similarly, the third one shows that  $t(z_0)$  makes an angle  $\theta(z_0)$  with the  $x$ -axis. It follows from (10.9) that

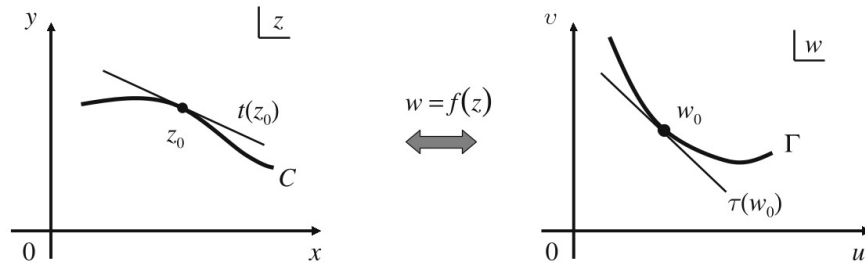
$$|\tau(w_0)| e^{i\psi(w_0)} = |f'(z_0)| |t(z_0)| e^{i[\phi(z_0) + \theta(z_0)]}.$$

Thus the magnitude of  $\tau(w_0)$  and its argument read

$$|\tau(w_0)| = |f'(z_0)| |t(z_0)| \quad \text{and} \quad \psi(w_0) = \phi(z_0) + \theta(z_0).$$

Each equation gives us the properties of the conformal mapping of a tangential line as follows:

- (i) The magnitude of the tangent  $|t(z_0)|$  is modified by the scale factor  $|f'(z_0)|$ , thus being enlarged or shrunk by the mapping. Since  $|f'(z_0)|$  depends on  $z_0$ , the magnification varies from point to point on  $C$ .



**Fig. 10.2.** Conformal mapping of a tangential line

- (ii) The angle between the tangent  $t(z)$  and the  $x$ -axis at  $z_0$  differs from the angle between the tangent  $\tau(w)$  and the  $u$ -axis at  $w_0$ . The difference is determined by the argument of  $f'(z_0)$ , denoted by  $\phi$ , called the argument of the mapping;  $\phi$  also depends on  $z_0$  and thus varies from point to point on  $C$ .

### 10.1.5 The Point at Infinity

For later use, we introduce a few concepts that are at the basis of further investigations on conformal mapping. Our aim is to understand the way in which the entire spherical curved surface is mapped conformally onto the entire flat plane with a one-on-one correspondence. This is achieved with the help of a stereographic projection between the complex plane and an artificial sphere as described below.

Let us consider a sphere of radius  $R$  (for convenience,  $R$  is taken as  $1/2$ ) such that the complex plane is tangential to it at the origin, as shown in Fig. 10.3. The point  $P$  on the sphere opposite the origin (called the north pole, for convenience) is used as the “eye” of the stereographic projection. We draw straight lines through  $P$  that intersect both the sphere and the plane. These lines permit a mapping of point  $z$  on the plane onto the point  $\zeta$  on the sphere (see Fig. 10.3). In this fashion the entire complex plane is mapped onto the sphere (called a **Riemann** or a **complex sphere**).

As to the properties of the Riemann sphere, the following statements can be verified without much difficulty.

1. Straight lines in the  $z$ -plane are mapped onto circles on the sphere that pass through  $P$ .

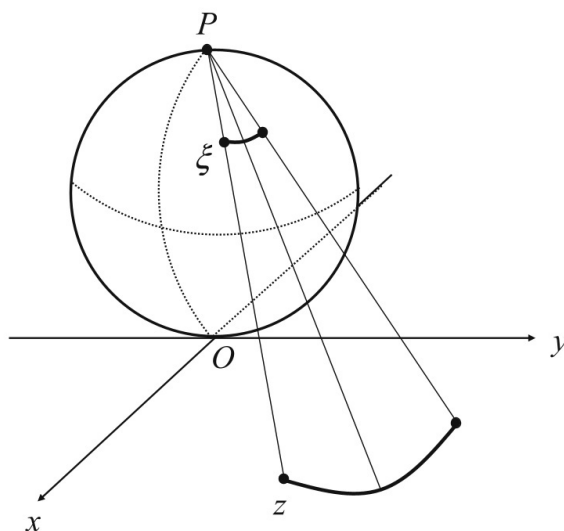


Fig. 10.3. Riemann sphere

2. The images of intersecting straight lines on the plane have two common points on the Riemann sphere, one of which is  $P$ .
3. The images of parallel straight lines on the  $z$ -plane have only the point  $P$  in common, and they have a common tangent at  $P$ .
4. The exterior of a circle  $|z| = R$  with  $R \gg 1$  is mapped onto the interior of a small spherical cap around point  $P$ . As  $R \rightarrow \infty$  the cap shrinks to  $P$ .

Note that the point  $P$  itself has no counterpart on the  $z$ -plane. Nevertheless, it has been found convenient to adjoin an extra point to the  $z$ -plane, known as the **point at infinity**, in such a way that a curve passing through  $P$  on the Riemann sphere is the image of a curve on the  $z$ -plane that approaches the point at infinity.

♠ **Point at infinity:**

The point at infinity  $z = \infty$  is defined as the point  $\tilde{z}$  that is mapped onto the origin  $z = 0$  by the transformation  $\tilde{z} = 1/z$ .

The importance of the point at infinity is greatly enhanced once we appreciate the conformal property of the stereographic projection: i.e., if two curves intersect on the  $z$ -plane at an angle  $\gamma$ , then their images on the sphere intersect at the same angle. This conformal property permits the definition of the angle between two *parallel* straight lines on the  $z$ -plane, i.e., the angle that their images make on the sphere at point  $P$ . (Indeed this angle is equal to zero as noted in **3** above.)

### 10.1.6 Singular Point at Infinity

The concept of a point at infinity is closely interwoven with the study of singularities of analytic functions. The notion of analyticity can be extended to a point at infinity by the following device: A function  $f(z)$  is considered to be analytic at infinity if the function

$$g(z) = f\left(\frac{1}{z}\right)$$

is analytic at  $z = 0$ . A more precise statement on this matter is given below.

♠ **Extended definition of conformal mappings:**

A function  $w = f(z)$  is said to transform the neighborhood of a point  $z_0$  conformally into a neighborhood of  $w = \infty$  if the function  $\eta = 1/f(z)$  transforms the neighborhood of  $z_0$  conformally into a neighborhood of  $\eta = 0$ .

*Example* The mapping  $w = 1/z$  is conformal at the origin  $z = 0$ . Initially, the function  $f(z) = 1/z$  is not defined at  $z = 0$ ; however, the subterfuge based on the Riemann sphere makes the mapping  $w = 1/z$  meaningful (and, furthermore, conformal) at  $z = 0$ . Note that it is also conformal at  $z = \infty$  even though the derivative  $f'(z)$  approaches zero as  $z \rightarrow \infty$ .

Owing to the above convention, it becomes possible to introduce the concept of a **pole at infinity**, a **branch at infinity**, and so on, through the corresponding behavior of  $g(z)$  at the origin. In fact, owing to our convention, a function  $f(z) = e^z$  that has no singularities in the original  $z$ -plane comes to possess an **essential singularity at infinity**. Other functions that have no singularities (e.g., all the polynomials in  $z$ ) are also found to have a breakdown of analyticity at infinity. In contrast, functions that are **analytic at infinity** possess at least one singularity for some finite value of  $z$ . The natural conjecture is that there may not be a perfectly analytic function. This problem has actually been resolved and is embodied in the theorem below.

♠ **Entire function:**

A function  $f(z)$  whose only singularity is an isolated singularity at the point at infinity  $z = \infty$  is called an **entire function** (or **integral function**). If this singularity is a pole of  $m$ th order, then  $f(z)$  must be a polynomial of degree  $m$ .

♠ **Liouville theorem:**

The only function  $f(z)$  that is analytic in the entire complex plane as well as at the point at infinity is the constant function  $f(z) = \text{const.}$

*Remark.* In some texts the term “complex plane” is tacitly assumed to mean the extended complex plane with the **point at infinity** included. Certain theorems may then be stated more conveniently. However, one should never forget that while there is a point at infinity, there is still no such thing as a complex number “infinity” in the sense that it possesses the algebraic properties shared by other complex numbers.

## Exercises

1. Suppose that two differential curves on the  $z$ -plane, meet at a point  $z_0$  at which  $f'(z_0) = f''(z_0) = \cdots = f^{(m-1)}(z_0) = 0$  and  $f^{(m)}(z_0) \neq 0$ . Show that the angle  $\theta$  between the two curves is magnified by  $m$  times through the conformal mapping  $w = f(z)$ .

**Solution:** From hypothesis,  $f(z)$  can be expanded in the neighborhood of the point  $z_0$  as

$$f(z) = f(z_0) + c_m(z - z_0)^m + c_{m+1}(z - z_0)^{m+1} + \cdots,$$

where  $c_m \neq 0$ . Then, by the same scenario as we used in deriving (10.1), the angle  $\tilde{\theta}$  between the mapped arcs at  $f(z_0)$  reads

$$\begin{aligned}\tilde{\theta} &= \lim_{\ell \rightarrow 0} \arg \frac{f(z_2) - f(z_0)}{f(z_1) - f(z_0)} = \lim_{\ell \rightarrow 0} \arg \left( \frac{z_2 - z_0}{z_1 - z_0} \right)^m \\ &= m \lim_{\ell \rightarrow 0} \arg \frac{z_2 - z_0}{z_1 - z_0} = m\theta. \quad \clubsuit\end{aligned}$$

2. We say that the mapping  $w = f(z)$  is **locally one-to-one** at  $z_0$  if  $f(z_1) \neq f(z_2)$  for any two distinct points  $z_1$  and  $z_2$  within the circle  $|z - z_0| < \delta$  with some  $\delta > 0$ . Show that  $w = f(z)$  is locally one-to-one at  $z_0$  if  $f(z)$  is analytic at  $z_0$  and  $f'(z_0) \neq 0$ .

**Solution:** Let  $f(z_0) = \alpha$  and take  $\delta > 0$  small enough so that  $f(z) - \alpha$  has no other zero in  $|z - z_0| < \delta$ . In view of the theorem regarding the isolated property of zeros, such a  $\delta$  can always be found. The argument principle says that

$$1 = \frac{1}{2\pi i} \oint_C \frac{f'(z)}{f(z) - \alpha} dz,$$

where  $C$  is a circle  $|z - z_0| = \delta$ . Denoting  $\Gamma = f(C)$ , we have

$$1 = \frac{1}{2\pi i} \oint_{\Gamma} \frac{dw}{w - \alpha} = \frac{1}{2\pi i} \oint_{\Gamma} \frac{dw}{w - \beta}$$

for any  $\beta$  satisfying  $|\beta - \alpha| < \varepsilon$  with sufficiently small  $\varepsilon$ . If we take  $\delta' \leq \delta$  so that

$$D = \{z; |z - z_0| < \delta'\} \subset f^{-1}[D^* = \{w; |w - \alpha| < \varepsilon\}],$$

it follows that for any  $z_1, z_2 \in D$ ,

$$1 = \frac{1}{2\pi i} \oint_{\Gamma} \frac{dw}{w - f(z_1)} = \frac{1}{2\pi i} \oint_{\Gamma} \frac{dw}{w - f(z_2)},$$

or equivalently,

$$1 = \frac{1}{2\pi i} \oint_{\Gamma} \frac{f'(z)}{f(z) - f(z_1)} dz = \frac{1}{2\pi i} \oint_{\Gamma} \frac{f'(z)}{f(z) - f(z_2)} dz.$$

This means that each function  $f(z) - f(z_1)$  and  $f(z) - f(z_2)$  has only one zero inside the circle  $|z - z_0| = \delta$ . Therefore, we conclude that  $f(z_1) \neq f(z_2)$  if  $z_1 \neq z_2$ .  $\clubsuit$

## 10.2 Elementary Transformations

### 10.2.1 Linear Transformations

The most simple conformal mapping  $w = f(z)$  would be the following:

♠ **Linear transformation:**

$$w = \alpha z + \beta, \quad (10.11)$$

where  $\alpha$  and  $\beta$  are complex numbers.

A linear transformation generates a translation plus a magnification and a rotation of a polygon, but does not affect its shape. Thus, for example, a line maps to a line, a rectangle maps to a rectangle, a circle maps to a circle, etc.

To appreciate the above statement, we first consider the particular case of  $\alpha = 1$ . From (10.11), we have

$$w = z + \beta, \quad (10.12)$$

which describes a translation by the constant  $\beta$  of the points being mapped. Obviously, a translation does not modify the length of a line or its orientation, only changes its position with respect to the coordinate axes. Since a polygon is constructed from three or more lines, the size and orientation of a polygon are not affected by a translation; only the position of the polygon is changed.

Next we consider the case of  $\beta = 0$ . When we express  $\alpha$  in polar form, the linear transformation becomes

$$w = |\alpha|e^{i\gamma}z$$

with a constant argument  $\gamma$ . Then, the line between two points transforms as

$$w_1 - w_2 = |\alpha|e^{i\gamma}(z_1 - z_2) = |\alpha| \cdot |z_1 - z_2|e^{i(\gamma+\theta)}.$$

Therefore, the length of a line in the  $z$ -plane,  $|z_1 - z_2|$ , becomes magnified by a constant factor  $|\alpha|$  and the line is rotated through an angle  $\gamma$ . Thus, the lengths of the sides of a polygon and the orientation of the polygon with respect to the axes is modified. Nevertheless, its shape remains unchanged by the linear transformation with  $\beta = 0$ .

We have seen that the values of  $\alpha$  and  $\beta$  straightforwardly determine the image of a polygon in the  $z$ -plane under a particular linear transformation. Conversely, if one knows the coordinates of two points on the original polygon in the  $z$ -plane and the images of those two points in the  $w$ -plane, one can determine  $\alpha$  and  $\beta$  and thus the linear transformation.

### 10.2.2 Bilinear Transformations

There is another important conformal mapping referred to as the **bilinear transformation** (or the **fractional** or **Möbius transformation**):

♠ **Bilinear transformation:**

$$w = \frac{\alpha z + \beta}{\gamma z + \delta}, \quad (10.13)$$

where  $\alpha, \beta, \gamma$  and  $\delta$  are complex numbers satisfying the relation  $\alpha\delta - \beta\gamma \neq 0$ .

The condition  $\alpha\delta - \beta\gamma \neq 0$  ensures that

$$\frac{df}{dz} = \frac{\alpha\delta - \beta\gamma}{(\gamma z + \delta)^2}$$

is nonzero at any finite point of the plane. Accordingly, the bilinear transformation (10.13) possesses the one-to-one property because if  $f(z_1) = f(z_2)$ , then

$$\frac{\alpha z_1 + \beta}{\gamma z_1 + \delta} = \frac{\alpha z_2 + \beta}{\gamma z_2 + \delta},$$

which implies  $(\alpha\delta - \beta\gamma)(z_1 - z_2) = 0$ , and thus  $z_1 = z_2$ .

*Remark.*

1. If  $\gamma = 0$ , the bilinear transformation (10.13) reduces to a linear transformation, which has already been discussed. Thus, we require that  $\gamma \neq 0$  in what follows.
2. The function  $f(z) = (\alpha z + \beta)/(\gamma z + \delta)$  serves as a **general solution** (see Sect. 15.1.4) of the differential equation:

$$\left(\frac{f''}{f'}\right)' - \frac{1}{2} \left(\frac{f''}{f'}\right)^2 = 0,$$

which is called the **Schwarz differential equation**.

Observe that the mapping (10.13) has two apparent exceptional points:  $z = \infty$  and  $z = -\delta/\gamma$  at which  $w$  diverges. It is possible to weed out these exceptions by extending the definition of conformal representation such that the point at infinity is included. With such an extension, the conformal property of the transformation (10.13) at the two points is recovered, even though the function  $f(z)$  itself diverges. Similarly, we can say that  $w = f(z)$  transforms the neighborhood of  $z = \infty$  conformally into that of a point  $w_0$  if  $w = \phi(\xi) = f(1/\xi)$  transforms the neighborhood of  $\xi = 0$  conformally into that of the point  $w_0$ .

A particularly interesting example of the bilinear transformation is

$$w = f(z) = \frac{z - z_0}{z - z_0^*}, \quad (10.14)$$

where  $\text{Im}(z_0) \neq 0$ . This transformation maps the upper half-plane of the  $z$ -plane including the  $x$ -axis, onto the unit circle centered at the origin of the  $w$ -plane. This is demonstrated in Exercise 1.

### 10.2.3 Miscellaneous Transformations

In what follows, we note several elementary transformations that facilitate a better understanding of the conformal nature of analytic functions. We shall see that any conformal transformation may be regarded as a transformation from Cartesian to **orthogonal curvilinear coordinates**.

*Example 1.*  $w = z^2$ ,  $w = \sqrt{z}$

Assume a conformal mapping defined by

$$w = z^2. \quad (10.15)$$

Setting  $z = x + iy$  and separating the real and imaginary parts, we have

$$x^2 - y^2 = u, \quad 2xy = v. \quad (10.16)$$

Thus, the straight lines parallel to the  $x$ - and  $y$ -axes in the  $z$ -plane denoted by

$$x = a \quad \text{and} \quad y = b$$

are mapped onto rectangular hyperbolas in the  $w$ -plane given by

$$u = a^2 - \frac{v^2}{4a^2} \quad \text{and} \quad u = \frac{v^2}{4b^2} - b^2,$$

respectively. This is shown schematically in Fig. 10.4.

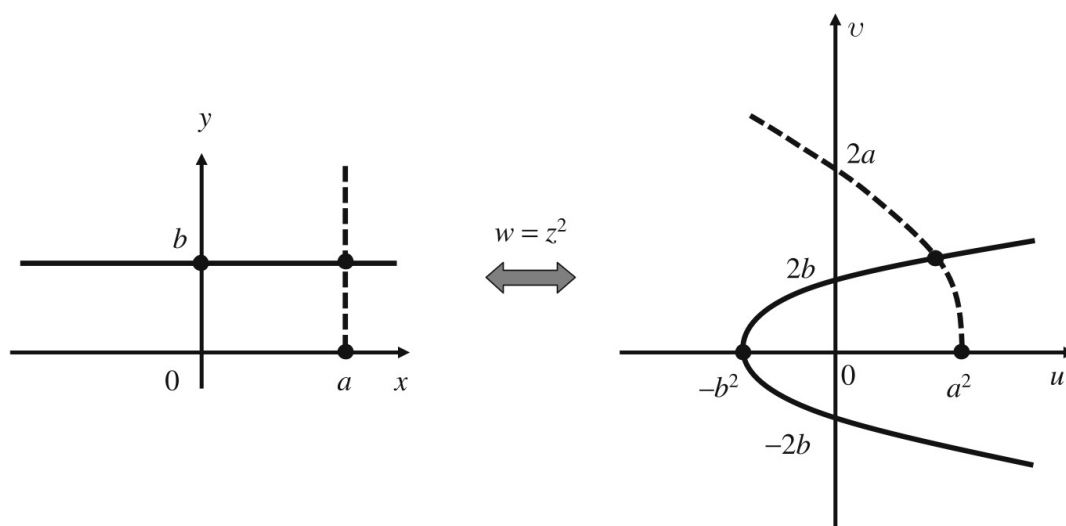
Another important feature of the mapping (10.15) is found by expressing  $z$  and  $w$  in polar coordinates:

$$z = \rho e^{i\phi}, \quad w = r e^{i\theta}.$$

On substitution in (10.15), we obtain

$$r = \rho^2, \quad \theta = 2\phi. \quad (10.17)$$

Hence, the upper half of the  $z$ -plane,  $0 \leq \phi \leq \pi$ , goes into the entire  $w$ -plane,  $0 \leq \theta \leq 2\pi$ ; the lower half also goes into the entire  $w$ -plane. In other words, points  $z$  and  $-z$  in the  $z$ -plane obviously go into the same point in the



**Fig. 10.4.** Mapping  $w = z^2$

$w$ -plane. This suggests the possibility that some distinct geometric figures in the  $z$ -plane may go into coincident figures in the  $w$ -plane.

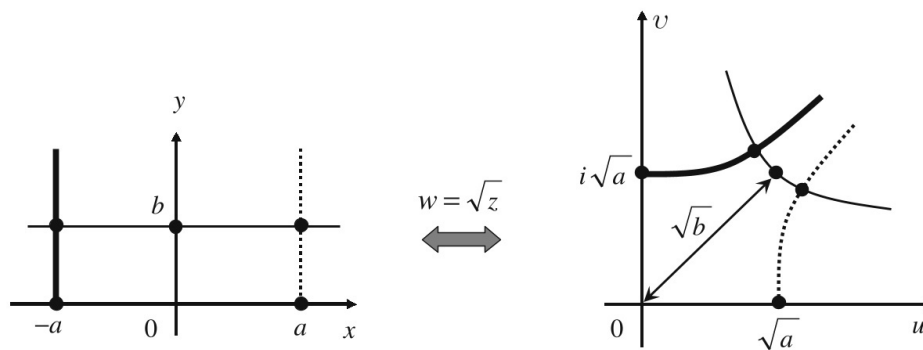
Next we consider the transformation:  $w = \sqrt{z}$ . In terms of polar forms, it reads

$$\sqrt{z} = \rho^{1/2} e^{i\phi/2} e^{in\pi},$$

so that we have

$$r = \sqrt{\rho}, \quad \theta = \frac{\phi}{2} + n\pi. \quad (10.18)$$

Owing to the additional term  $n\pi$  in the latter equation in (10.18), a half revolution in the  $z$ -plane corresponds to one complete revolution in the  $w$ -plane. This is obviously a manifestation of the multivaluedness of the root function. The mapping of the upper half of the  $z$ -plane onto the  $w$ -plane is illustrated schematically in Fig. 10.5.



**Fig. 10.5.** Mapping  $w = \sqrt{z}$

*Example 2.*  $w = e^z$ ,  $w = \log z$

In the case of

$$w = e^z, \quad (10.19)$$

there are simple relationships between the Cartesian coordinates in the  $z$ -plane and the polar coordinates in the  $w$ -plane

$$re^{i\theta} = e^{x+iy} = e^x(\cos y + i \sin y); \text{ i.e., } r = e^x, \theta = y.$$

The lines  $x = \text{const.}$ , parallel to the  $y$ -axis, become concentric circles in the  $w$ -plane; the lines  $y = \text{const.}$ , parallel to the  $x$ -axis, become rays emerging from the origin. Accordingly, a strip of the  $z$ -plane bounded by  $y = y_0$  and  $y = y_0 + 2\pi$  goes into the entire  $w$ -plane.

In the inverse of (10.19)

$$z = \log w, \quad x = \log r, \quad y = \theta + 2n\pi,$$

which is an infinitely many-valued function since all points for different values of  $n$  correspond to the same point in the  $w$ -plane.

*Example 3.*  $w = \cosh z$

Next let us consider the following functions:

$$w = \cosh z.$$

The Cartesian coordinates in the two planes are related as follows:

$$\begin{aligned} u + iv &= \cosh(x + iy) = \cosh x \cos y + i \sinh x \sin y, \\ u &= \cosh x \cos y, \quad v = \sinh x \sin y. \end{aligned} \quad (10.20)$$

Dividing the first equation by  $\cosh x$ , the second by  $\sinh x$ , squaring and adding, we have an ellipse in the  $w$ -plane that corresponds to the straight line  $x = \text{const.}$  in the  $z$ -plane. Similarly,  $y = \text{const.}$  goes into a hyperbola in the  $w$ -plane. The equations of the ellipses and hyperbolas are

$$\frac{u^2}{\cosh^2 x} + \frac{v^2}{\sinh^2 x} = 1, \quad \frac{u^2}{\cos^2 y} - \frac{v^2}{\sin^2 y} = 1. \quad (10.21)$$

The semimajor and semiminor axes of the ellipses are  $\cosh x$  and  $\sinh x$ ; the semifocal distance is unity. The semiaxes of the hyperbolas are  $\cos y$  and  $\sin y$ ; the semifocal distance is unity. Hence, equations (10.21) represent families of confocal ellipses and hyperbolas. This transformation may be regarded as a transformation from Cartesian to elliptic coordinates.

*Example 4.*  $w = 1/z$

Consider the function

$$w = \frac{1}{z} \quad (10.22)$$

and use rectangular coordinates to obtain

$$(u + iv)(x + iy) = 1.$$

By equating real and imaginary parts, we set

$$ux - vy = 1, \quad vx + uy = 0.$$

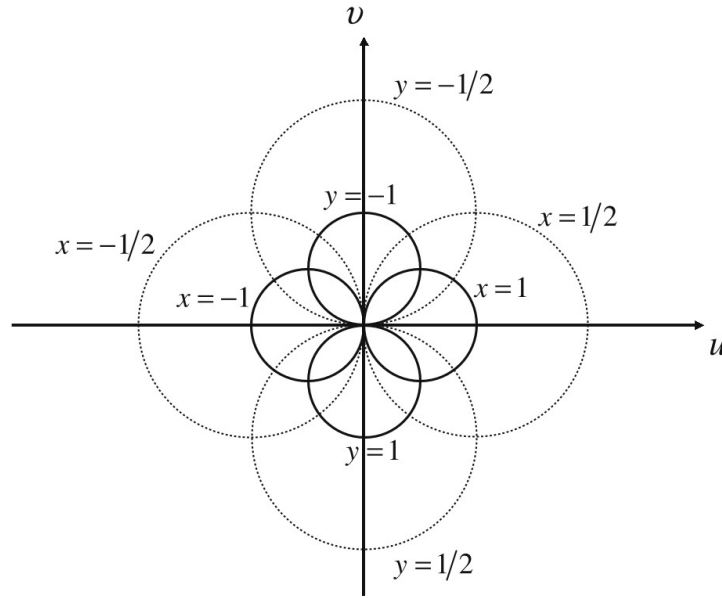
By an algebraic elimination first of  $x$  and then of  $y$ , we arrive at the two families of circles:

$$u^2 + \left(v + \frac{1}{2y}\right)^2 = \frac{1}{4y^2}, \quad \left(u - \frac{1}{2x}\right)^2 + v^2 = \frac{1}{4x^2}. \quad (10.23)$$

The degenerate cases  $x = 0$  and  $y = 0$  cannot be handled by (10.23), but from (10.22) we find that respectively, they give the two axes  $u = 0$  and  $v = 0$ .

The transformation is shown in Fig. 10.6. Note that through the transformation, the edge of the  $z$ -plane at infinity ( $z = \infty$ ) is pulled into the origin of the  $w$ -plane ( $w = 0$ ), whereas the center of the  $z$ -plane is stretched out in all directions to infinity in the  $w$ -plane. It is possible to visualize this process by introducing an artificial concept, called “the **point at infinity**”; see Sect. 10.1.5 for details.

**Remark.** The mapping  $w = 1/z$  reverses the orientation of the circumference of the circle to be mapped:  $\arg(w) = -\arg(z)$ . For example, the circumference of  $|w| = 1$  is described in the negative sense since if  $|z| = 1$  is described in the positive sense.



**Fig. 10.6.** Mapping  $w = 1/z$

### 10.2.4 Mapping of Finite-Radius Circle

Remember that the **analyticity** of functions is characterized by the isotropy of their derivatives. Owing to the isotropy, infinitely small circles on the  $z$ -plane are transformed into infinitely small circles on the  $w$ -plane through any analytic function  $w = f(z)$ . Of course, this shape-preserving behavior disappears when the circle has a finite radius; because the scale factor  $h$  generally depends on  $z$ . Nevertheless, there exist a class of nontrivial analytic functions that transform a finite circle on the  $z$ -plane onto the  $w$ -plane, which is simply a bilinear transformation.

♠ **Theorem:**

Bilinear transformations  $w = f(z)$  map circles (or straight lines) on the  $z$ -plane onto circles (or straight lines) on the  $w$ -plane.

*Proof* Our proof is based on the fact that the bilinear transformation formula (10.11) can be rewritten as

$$w = f(z) = \frac{\alpha}{\gamma} + \frac{\beta\gamma - \alpha\delta}{\gamma} \frac{1}{\gamma z + \delta}.$$

This is composed of a sequential transformation of the following:

1.  $w = z + b$ , a simple translation of the plane by the complex vector  $b$ .
2.  $w = az$ , a rotation of the plane through the angle  $\arg a$ , followed by an expansion (or contraction) by  $|a|$ .
3.  $w = 1/z$ , an **inversion** that takes the interior of the unit circle to the exterior and vice versa.

Since these transformations are all conformal, their composition surely maps circles (or straight lines) onto circles (or straight lines). ♣

*Remark.* Statement **3** above regarding the inversion  $w = 1/z$  is followed by considering the equation

$$\alpha(x^2 + y^2) + \beta x + \gamma y + \delta = 0,$$

which represents a circle ( $\alpha \neq 0$ ) or straight line ( $\alpha = 0$ ) in the  $z$ -plane. This can be written as

$$\alpha|z|^2 + \frac{\beta}{2}(z + z^*) + \frac{\gamma}{2i}(z - z^*) + \delta = 0. \quad (10.24)$$

Then, the transformation  $w = 1/z$  maps it onto

$$\delta|w|^2 + \frac{\beta}{2}(w + w^*) - \frac{\gamma}{2i}(w - w^*) + \alpha = 0,$$

which is a circle ( $\delta \neq 0$ ) or a straight line ( $\delta = 0$ ).

### 10.2.5 Invariance of the Cross ratio

The following peculiarity of a Möbius transformation serves as a useful device in applications of conformal mapping.

♠ **Invariance of the cross ratio:**

Any Möbius transformation  $w = f(z)$  that maps the four points  $z_i$  ( $i = 1, 2, 3, 4$ ) into  $w_i$  ( $i = 1, 2, 3, 4$ ), respectively, satisfies

$$\frac{(w_1 - w_4)(w_3 - w_2)}{(w_1 - w_2)(w_3 - w_4)} = \frac{(z_1 - z_4)(z_3 - z_2)}{(z_1 - z_2)(z_3 - z_4)} \equiv \lambda.$$

The constant  $\lambda$  is called the **cross ratio** (or **anharmonic ratio**).

*Proof* Let  $z_i$  ( $i = 1, 2, 3, 4$ ) be four distinct finite points on the  $z$ -plane and let  $w_i$  ( $i = 1, 2, 3, 4$ ) be their corresponding images through a Möbius transformation. Then, for any two of the points, we have

$$w_k - w_i = \frac{\alpha z_k + \beta}{\gamma z_k + \delta} - \frac{\alpha z_i + \beta}{\gamma z_i + \delta} = \frac{\alpha\delta - \beta\gamma}{(\gamma z_k + \delta)(\gamma z_i + \delta)}(z_k - z_i),$$

and, consequently, for all four,

$$\frac{(w_1 - w_4)(w_3 - w_2)}{(w_1 - w_2)(w_3 - w_4)} = \frac{(z_1 - z_4)(z_3 - z_2)}{(z_1 - z_2)(z_3 - z_4)}. \quad (10.25)$$

This clearly ensures the invariance of the cross ratio  $\lambda$  under the Möbius transformation. ♣

*Remark.* If one of the points of  $w_i$ , say  $w_1$ , is the **point at infinity**, the corresponding result is obtained by letting  $w_1 \rightarrow \infty$  in (10.25). The left-hand side then takes the form

$$\frac{w_3 - w_2}{w_3 - w_4}.$$

This expression is to be regarded as the cross ratio of the points  $\infty, w_2, w_3, w_4$ . A similar remark applies if one of the points  $z_i$  is the point at infinity.

If  $z_4$  is taken to be a variable  $z$ , then the corresponding image  $w_4$  on the  $w$ -plane becomes a function of  $z$  that obeys the relation

$$\frac{(w_1 - w)(w_3 - w_2)}{(w_1 - w_2)(w_3 - w)} = \frac{(z_1 - z)(z_3 - z_2)}{(z_1 - z_2)(z_3 - z)}. \quad (10.26)$$

By solving (10.26) for  $w$ , we can verify that it transforms the three points  $z_1, z_2, z_3$  into the corresponding points  $w_1, w_2, w_3$ . In this context, the expression (10.26) turns out to show that a Möbius transformation is uniquely

determined by three correspondences. Since a circle is uniquely determined by three points on its circumference, (10.26) can be used to find Möbius transformations that map a given circle determined by  $z_i (i = 1, 2, 3)$  onto a second given circle (or straight line) determined by  $w_i (i = 1, 2, 3)$ .

*Example* If we take  $z_1 = 1, z_2 = i, z_3 = -1$  and  $w_1 = 0, w_2 = 1, w_3 = \infty$ , we obtain the transformation

$$w = i \frac{1 - z}{1 + z}.$$

This maps the circle  $|z| = 1$  on the real axis and the interior  $|z| < 1$  of the unit circle on the upper half of the  $w$ -plane.

## Exercises

1. Consider the function  $w = f(z) = (z - z_0)/(z - z_0^*)$  with  $\text{Im}(z_0) \neq 0$ . Show that it maps the region  $\text{Im}z > 0$  onto  $|w| < 1$ .

**Solution:** Set  $z = x$  to obtain

$$|w|^2 = \left( \frac{x - z_0}{x - z_0^*} \right) \left( \frac{x - z_0}{x - z_0^*} \right)^* = \left( \frac{x - z_0}{x - z_0^*} \right) \left( \frac{x - z_0^*}{x - z_0} \right) = 1.$$

That is, the image on the  $x$ -axis is the circumference of the unit circle centered at the origin of the  $w$ -plane.

Next we evaluate the image of a point off the  $x$ -axis in the upper half of the  $z$ -plane. Expressing  $z$  and  $z_0$  in polar form, we have

$$|w|^2 = \frac{(re^{i\theta} - r_0e^{i\theta_0})(re^{-i\theta} - r_0e^{-i\theta_0})}{(re^{i\theta} - r_0e^{-i\theta_0})(re^{-i\theta} - r_0e^{i\theta_0})} = \frac{\xi_1 - \xi_2}{\xi_1 + \xi_2}, \quad (10.27)$$

where

$$\xi_1 = r^2 + r_0^2 - 2rr_0 \cos \theta \cos \theta_0 \quad \text{and} \quad \xi_2 = 2rr_0 \sin \theta \sin \theta_0.$$

Since  $-1 \leq \cos \theta \cos \theta_0 < 1$ , we have

$$(r - r_0)^2 \leq r^2 + r_0^2 - 2rr_0 \cos \theta \cos \theta_0 = \xi_1, \quad \text{i.e., } \xi_1 \geq 0.$$

In addition, since  $z$  and  $z_0$  are in the upper half-plane, both  $\sin \theta$  and  $\sin \theta_0$  are positive, so  $\xi_2 > 0$ . Consequently, we have

$$|w|^2 < 1,$$

which means that the images of points in the upper half of the  $z$ -plane are located in the interior of the unit origin-centered circle. ♣

**Remark.** If  $z_0$  were real, all points  $z$  would be mapped onto the single point  $w = 1$ , which is the reason we assumed  $\text{Im}(z_0) \neq 0$  in the first place.

2. Show that  $w = (z - z_0)/(z_0^*z - 1)$  in which  $|z_0| < 1$  maps  $|z| < 1$  onto  $|w| < 1$  and  $z = z_0$  onto  $w = 0$ .

**Solution:** Observe that

$$\begin{aligned} 1 - |w|^2 &= 1 - \frac{|z - z_0|^2}{|z_0^*z - 1|^2} = \frac{|z_0|^2|z|^2 - z^2 - (z_0^*)^2 + 1}{|z_0^*z - 1|^2} \\ &= \frac{(1 - |z|^2)(1 - |z_0|^2)}{|z_0^*z - 1|^2}. \end{aligned}$$

Hence,  $|z| = 1$  corresponds to  $|w| = 1$ . In addition,  $z = z_0$  corresponds to  $w = 0$ . These mean that  $|z| < 1$  is transformed onto  $|w| < 1$ . ♣

3. Let  $C$  and  $C^*$  be two simple closed contours in the  $z$ - and the  $w$ -plane, respectively, and let  $w = f(z)$  be analytic within and on  $C$ . If  $w = f(z)$  maps  $C$  onto  $C^*$  in such a way that  $C^*$  is traversed by  $w$  exactly once in the positive sense under the condition that  $z$  describes  $C$  in the positive sense, then  $w = f(z)$  maps the domain bounded by  $C$  onto the domain bounded by  $C^*$ .

**Solution:** We denote the domains bounded by  $C$  and  $C^*$  by  $D$  and  $D^*$ , respectively. Then it suffices to prove that every point of  $D^*$  is taken exactly once if  $z$  is in  $D$ . Recall that the number  $n$  of zeros of the function  $w_0 - f(z)$  in  $D$  is given by

$$n = \frac{1}{2\pi i} \oint_C \frac{f'(z)}{f(z) - w_0} dz.$$

With the substitution  $w = f(z)$ ,  $f'(z)dz = dw$ , this is rewritten as

$$n = \frac{1}{2\pi i} \oint_{C^*} \frac{dw}{w - w_0},$$

where the integration has to be extended over the contour  $C^*$  into which  $C$  is transformed by  $w = f(z)$ . By the residue theorem, the value of this expression is 1 if  $w_0$  is within  $C^*$  and 0 if  $w_0$  is outside  $C^*$ . This shows that every point in  $D^*$  is taken exactly once and that a value outside  $D^*$  is not taken at all. This completes the proof. ♣

4. Find a conformal mapping  $w = f(z)$  of the region between the two circles  $|z| = 1$  and  $|z - (1/4)| = 1/4$  onto an annulus  $a < |z| < 1$ .

**Solution:** To solve this, we have to find a bilinear transformation that simultaneously maps  $|z| < 1$  onto  $|z| < 1$  and  $|z - (1/4)| < 1/4$  onto a disc of the form  $|z| < a$ . Note that

$$w = \frac{z - \alpha}{1 - \alpha^* z}$$

maps  $|z| < 1$  onto  $|z| < 1$ , and that

$$g(z) = a \frac{4z - 1 - \beta}{1 - \beta^*(4z - 1)}$$

maps  $|z| < 1$  and  $|z - (1/4)| < 1/4$  onto a disc of the form  $|z| < a$ .

Equating coefficients leads us to  $\alpha = 2 - \sqrt{3}$ . ♣

5. Find the bilinear transformation that maps  $z = 0, i, -1$  onto  $w = 1, -1, 0$ , respectively.

**Solution:** Set  $[z, 0, i, -1] = [w, 1, -1, 0]$  to obtain  $w = -(z + i)/(3z - i)$ . ♣

6. Show that four distinct arbitrary points on the  $z$ -plane can be mapped through the bilinear transformation onto  $w = 1, -1, c, -c$  on the  $w$ -plane, where  $c$  is a complex number depending on the cross ratio  $\lambda$  of the mapping. Determine an explicit form of  $c$  as a function of  $\lambda$ .

**Solution:** Let  $[z_1, z_2, z_3, z_4] = [1, -1, c, -c]$  to obtain  $c = (1 + \lambda \pm 2\sqrt{2})/(1 - \lambda)$  and  $c_1 c_2 = 1$ . ♣

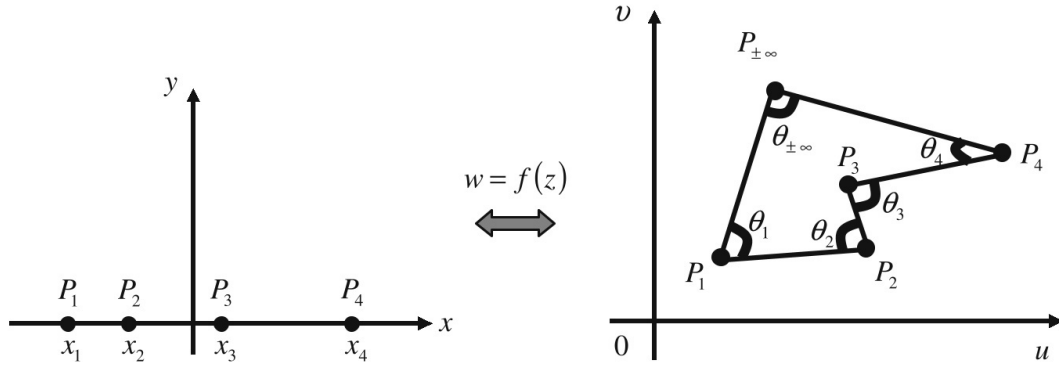
## 10.3 Applications to Boundary-Value Problems

### 10.3.1 Schwarz–Christoffel Transformation

In the preceding section, we discussed rich properties of the bilinear transformation that can transform the upper half of the  $z$ -plane onto the unit circle of the  $w$ -plane. Now we turn to a similar kind of important mappings called the **Schwarz–Christoffel transformation** (abbreviated SC transformation), which transforms the upper (or lower) half of the  $z$ -plane onto the inside of a  $n$ -sided polygon drawn on the  $w$ -plane. This transformation is defined by the following integral:

$$w(z) = \beta + \alpha \sum_{i=1}^n \int_{x_i}^z (z' - x_i)^{-\theta_i/\pi} dz'. \quad (10.28)$$

Here  $x_i$  ( $1 \leq i \leq n$ ) are  $n$  distinct fixed points along the  $x$ -axis, and the angle  $\theta_j$  is defined as shown in Fig. 10.7, being either positive or negative according to whether we follow the boundary of the polygon counterclockwise or clockwise. (For example,  $\theta_1$  and  $\theta_2$  are positive, but  $\theta_3$  is negative in Fig. 10.7.) The constant  $\alpha$  gives rise to a magnification of that image by a factor  $|\alpha|$  and a rotation of that image by an angle  $\arg(\alpha)$ . The constant  $\beta$  generates a translation of the magnified and rotated image.



**Fig. 10.7.** Schwarz–Christoffel transformation of the real axis of the  $z$ -plane to a polygon on the  $w$ -plane

*Remark.* If we wish to transform the upper half of the  $z$ -plane into the exterior of the polygon in the  $w$ -plane, it suffices to define

$$w(z) = \beta + \alpha \int^z (z' - x_1)^{\theta_1/\pi} (z' - x_2)^{\theta_2/\pi} \cdots (z' - x_N)^{\theta_N/\pi} dz',$$

where the  $\theta$ 's are assigned the same values as in the preceding case.

*Example* The function

$$w = f(z) = \int^z \frac{d\xi}{\sqrt{(1 - \xi^2)(1 - k^2 \xi^2)}} \quad (0 < k < 1) \quad (10.29)$$

maps the upper half of the  $z$ -plane ( $\text{Im}z > 0$ ) into the interior of a rectangle on the  $w$ -plane. In fact, (10.29) is obtained by putting  $n = 4$ ,  $\theta_i = \pi/2$  for all  $i = 1, 2, 3, 4$  in the definition (10.28), followed by setting  $x_1 = 1$ ,  $x_2 = -1$ ,  $x_3 = 1/k$  and  $x_4 = -1/k$ , all of which are located on the real axis. The integral in (10.29) is called an **elliptic integral of the first kind**.

### 10.3.2 Derivation of the Schwartz–Christoffel Transformation

In order to derive equation (10.28) for the Schwarz–Christoffel transformation, we let

$$x_1 < x_2 < \cdots < x_n$$

be points on the real axis and consider the function  $f(z)$  whose derivative is

$$f'(z) = \alpha(z - x_1)^{-k_1} (z - x_2)^{-k_2} \cdots (z - x_n)^{-k_n}. \quad (10.30)$$

For this function we have

$$\arg f'(z) = \arg \alpha - k_1 \arg(z - x_1) - k_2 \arg(z - x_2) - \cdots - k_n \arg(z - x_n).$$

Now, visualize the point  $z$  as moving from left to right along the real axis, starting to the left of the point  $x_1$ . When  $z < x_1$ , we have

$$\arg(z - x_1) = \arg(z - x_2) = \cdots = \arg(z - x_n) = \pi,$$

whereas for  $x_1 < z < x_2$ ,  $\arg(z - x_1) = 0$ , the others remaining at  $\pi$ . Hence, as  $z$  crosses  $a_1$  from left to right,  $\arg f'(z)$  increases by  $k_1\pi$ . It remains constant for  $x_1 < z < x_2$  and increases by  $k_2\pi$  as  $z$  crosses  $x_2$ , etc. As a result, the image of the segment  $-\infty < z < a_1$  becomes a straight line, the image of  $x_1 < z < x_2$  being another whose argument exceeds that of the first by  $k_1\pi$ , and so on.

If we constrain the numbers  $k_1, \dots, k_n$  to lie between  $-1$  and  $1$ , then the increments in the argument of  $f'(z)$  will lie between  $-\pi$  and  $\pi$ . Further, for  $k_1 < 1, k_2 < 1, \dots, k_n < 1$ , it is obvious that the function  $f(z)$  whose derivative is (10.30) is actually continuous at each of the points  $x_1, x_2, \dots, x_n$ . Therefore, the image of the moving point  $z$  will be a polygonal line. Finally, integrate (10.30) to set the equation

$$f(z) = \beta + \alpha \int^z (z' - a_1)^{-k_1} (z' - a_2)^{-k_2} \cdots (z' - a_n)^{-k_n} dz', \quad (10.31)$$

which maps the  $x$ -axis onto a polygonal line.

*Remark.*

1. The sum of the exterior angles of this polygonal line is

$$k_1\pi + k_2\pi + \cdots + k_n\pi = \pi \sum_{i=1}^n k_i.$$

Hence, in order for the polygon to be closed, it is necessary that  $\sum_{i=1}^n k_i = 2$ . Particularly when  $k_i > 0$  for all  $i$ , then the polygon becomes convex.

2. The complex constants,  $\alpha$  and  $\beta$ , control the position, size, and orientation of the polygon. Thus  $\beta$  may be so chosen that one of the vertices of the polygon will coincide with some specified point e.g., the origin. Then  $\alpha$  may be chosen so that one side of the polygon will be of given size and parallel to a given direction.

### 10.3.3 The Method of Inversion

The Schwarz–Christoffel transformation itself is applicable to polygons composed of straight lines, but not to those of circular ones. Nevertheless, combining the **method of inversion**, the former transformation can be extended to regions bounded by circular arcs.

♠ **Inversion with respect to a circle:**

An inversion transformation  $w = f(z)$  with respect to a circle  $|z| = a$  is defined by

$$w = \frac{a^2}{z^*}. \quad (10.32)$$

through which the interior points of the circle are mapped onto exterior points, and vice versa.

The inversion preserves the magnitude of the angle between two intersecting curves, but it reverses the sign of the angle. This is attributed to the fact that (10.32) consists of two successive transformations: the first  $a^2/z$ , and the second a reflection with respect to the real axis. The first of these is conformal, whereas the second maintains the angle but reverses its sign.

For the purpose of this section, we investigate the inversion of a circle of radius  $|z_0|$  centered at  $z = z_0 \neq 0$ . This circle is expressed by

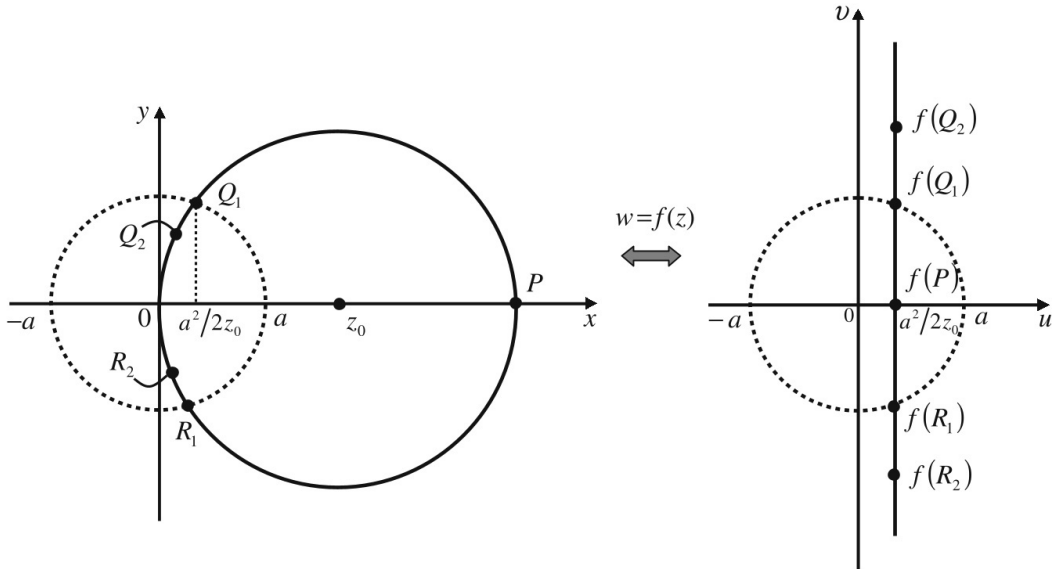
$$|z - z_0| = |z_0| \quad (10.33)$$

or

$$z^*z - z_0(z + z^*) = 0. \quad (10.34)$$

Note that this circle passes through the origin, i.e., the center of an inversion circle. Through the inversion (10.32), the circle (10.34) is mapped onto

$$\frac{a^4}{ww^*} - z_0 \left( \frac{a^2}{w} + \frac{a^2}{w^*} \right) = 0.$$



**Fig. 10.8.** Inversion of the circle  $|z - z_0| = |z_0|$  in (10.33) with respect to a circle  $|z| = a$  through the mapping  $w = a^2/z^*$  given in (10.32)

By multiplying  $ww^*$  on both sides and putting  $w = u + iv$ , we have

$$a^4 - 2a^2 z_0 u = 0,$$

or equivalently,

$$u = \frac{a^2}{2z_0}.$$

This means that by the inversion, the circle (10.33) is mapped onto a straight line parallel to the imaginary axis of the  $w$ -plane (see Fig. 10.8).

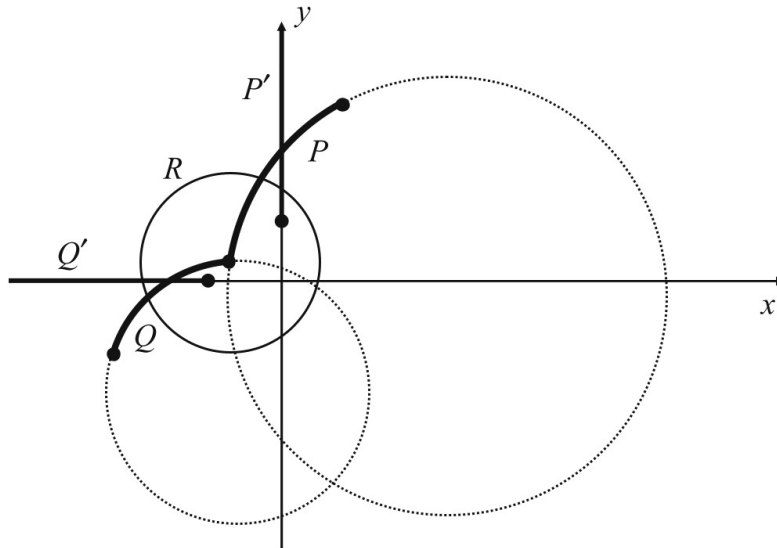
The role that inversion plays in extending the Schwarz–Christoffel transformation should now be clear. Assume two interesting circular arcs such as  $P$  and  $Q$  in Fig. 10.9 and a circle  $R$  of radius  $a$  whose center is the intersection of the two circular arcs. Then, by an inversion with respect to  $R$ , the point at the intersection is transformed into the **point at infinity**, the arcs themselves being transformed into the solid portions of the lines  $P'$  and  $Q'$ . As a result, the Schwarz–Christoffel transformation may now be applied to these two straight lines, whereas it may not be applied to the original circular arcs.

## Exercises

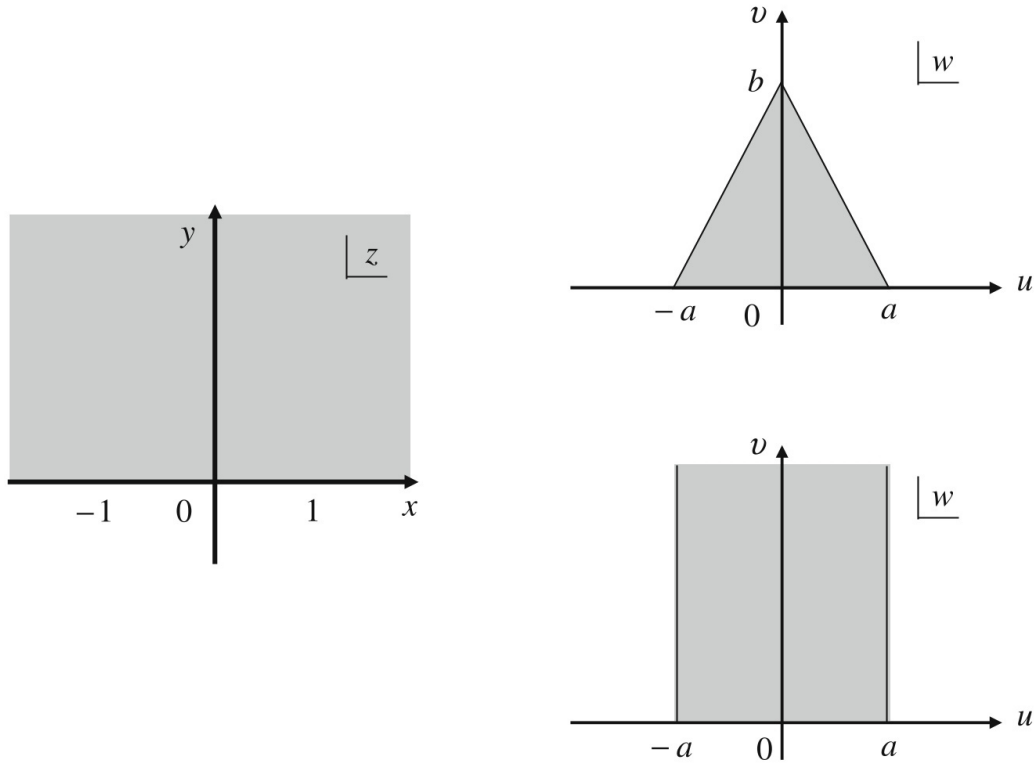
1. Find a transformation that maps the upper half of the  $z$ -plane onto the triangular region shown in Fig. 10.10 in such a way that the points  $x_1 = -1$  and  $x_2 = 1$  are mapped onto the points  $w = -a$  and  $w = a$ , respectively, and the point  $x_3 = \pm\infty$  is mapped onto  $w = ib$ .

**Solution:** Let us denote the angles at  $w_1$  and  $w_2$  in the  $w$ -plane by  $\phi_1 = \phi_2 = \phi$ , where  $\phi = \tan^{-1}(b/a)$ . Since  $x_3$  is taken at infinity we may omit the corresponding factor in (10.28) to obtain

$$w = \beta + \alpha \int_0^z (\xi + 1)^{-\phi/\pi} (\xi - 1)^{-\phi/\pi} d\xi = \beta + \alpha \int_0^z (\xi^2 - 1)^{-\phi/\pi} d\xi. \quad (10.35)$$



**Fig. 10.9.** Inversion of circular arcs  $P$  and  $Q$  with respect to the circle  $R$



**Fig. 10.10.** Mapping of the upper half of the  $z$ -plane onto a certain limited region of the  $w$ -plane

The required transformation may then be found by fixing the constants  $\alpha$  and  $\beta$  as follows. Since the point  $z = 0$  lies on the line segment  $x_1x_2$  it will be mapped onto the line segment  $w_1w_2$  in the  $w$ -plane, and by symmetry must be mapped onto the point  $w = 0$ . Thus setting  $z = 0$  and  $w = 0$  in (10.35), we obtain  $\beta = 0$ . An expression for  $\alpha$  can be found by considering the region in the  $w$ -plane in Fig. 10.10 to be the limiting case of the triangular region with the vertex  $w_3$  at infinity. Thus we may use the above, but with the angles at  $w_1$  and  $w_2$  set to  $\phi = \pi/2$ . From (10.35), we obtain  $w = \alpha \int_0^z (1/\sqrt{\xi^2 - 1}) d\xi = i\alpha \sin^{-1} z$ . By setting  $z = 1$  and  $w = a$ , we find  $i\alpha = 2a/\pi$ , so the required transformation is  $w = (2a/\pi) \sin^{-1} z$ . ♣

- 2.** Find the conformal mapping that transforms the interior of the circle  $|z| < 1$  to the interior of a polygon on the  $w$ -plane, subject to the condition that the points  $z_1, z_2, \dots, z_n$  lying on the circle  $|z| = 1$  are mapped, respectively, onto the vertex  $w_1, w_2, \dots, w_n$  of the polygon.

**Solution:** Consider first the transformation

$$\tau(z) = \frac{-iz + 1}{z - i}, \quad (10.36)$$

which maps  $|z| < 1$  onto  $\text{Im} \tau > 0$ . It yields

$$\frac{d\tau}{dz} = -\frac{2}{(z-i)^2} \quad \text{and} \quad \tau - \tau_j = \frac{2(z_j - z)}{(z-i)(z_j - i)} \quad (j = 1, 2, \dots, n). \quad (10.37)$$

Next, we assume that through (10.36), the points  $z_1, z_2, \dots, z_n$  are mapped, respectively, onto the points  $\tau_1, \tau_2, \dots, \tau_n$  that are located on the line  $\text{Im}\tau = 0$ . Then, the transformation that maps  $\text{Im}\tau > 0$  onto the interior of a polygon on the  $w$ -plane is given by  $w = w(\tau)$ , whose derivative reads

$$\frac{dw}{d\tau} = \alpha(\tau - \tau_1)^{(k_1/\pi)-1}(\tau - \tau_2)^{(k_2/\pi)-1} \dots (\tau - \tau_n)^{(k_n/\pi)-1}. \quad (10.38)$$

Here  $k_i$  is the internal angle of the polygon at the  $i$ th vertex, which satisfies  $\sum_{i=1}^n k_i = (n-2)\pi$ . From (10.37) and (10.38), we have

$$\frac{dw}{dz} = -\frac{2\alpha}{(z-i)^2} \cdot \frac{1}{2^2} \cdot \frac{(z_1 - z)^{(k_1/\pi)-1} \dots (z_n - z)^{(k_n/\pi)-1}}{(z-i)^{-2}(z_1 - i)^{(k_1/\pi)-1} \dots (z_n - i)^{(k_n/\pi)-1}}.$$

Replace  $(\alpha/2)(z_1 - i)^{1-(k_1/\pi)} \dots (z_n - i)^{1-(k_n/\pi)}$  by  $\alpha$  to obtain the final result:

$$w = f(z) = \alpha \int_{z_0}^z (z_1 - \zeta)^{(k_1/\pi)-1} \dots (z_n - \zeta)^{(k_n/\pi)-1} d\zeta + \beta,$$

where  $\alpha (\neq 0), \beta$  are complex constants and  $z_0 \neq z_1, \dots, z_n$ . ♣

**3.** Prove that the function

$$w = f(z) = \int_0^z \frac{1}{\sqrt[3]{1 - \xi^6}} d\xi \quad (10.39)$$

maps the unit circle on the  $z$ -plane onto a regular hexagon on the  $w$ -plane.

**Solution:** Observe that  $\xi^6 - 1 = (\xi - \xi_1) \dots (\xi - \xi_6)$  with  $|\xi_j| = 1$  ( $j = 1, \dots, 6$ ). Similarly to Exercise 2 above, we map  $|z| < 1$  onto  $\text{Im}\tau > 0$ , and then let the points  $\tau_1, \dots, \tau_6$  located on the line  $\text{Im}\tau = 0$  correspond to  $\xi_1, \dots, \xi_6$ . Then, by setting  $n = 6$  and  $k_j = (2/3)\pi$  for all  $j$ , we see that the transformation (10.39) maps  $|z| < 1$  onto a regular hexagon on the  $w$ -plane,  $(\sqrt[3]{2}/6)\Gamma(1/3)$  on a side.

**4.** Suppose that  $\phi(z)$  satisfies the Laplace equation and let  $w = f(z)$  be a conformal mapping. Then, show that the function

$$\phi(w) = \phi(u, v)$$

also satisfies Laplace's equation in the  $w$ -plane; i.e.,

$$\frac{\partial^2 \phi}{\partial u^2} + \frac{\partial^2 \phi}{\partial v^2} = 0. \quad (10.40)$$

**Solution:** Since  $x = x(u, v)$ , the partial derivative  $\partial/\partial x$  can be rewritten as  $\partial/\partial x = u_x(\partial/\partial u) + v_x(\partial/\partial v)$ , where  $u_x = \partial u/\partial x$  and  $v_x = \partial v/\partial x$ . It yields

$$\begin{aligned}\frac{\partial^2 \phi}{\partial x^2} &= \left( u_x \frac{\partial}{\partial u} + v_x \frac{\partial}{\partial v} \right) \left( u_x \frac{\partial}{\partial u} + v_x \frac{\partial}{\partial v} \right) \phi \\ &= (u_x)^2 \frac{\partial^2 \phi}{\partial u^2} + (v_x)^2 \frac{\partial^2 \phi}{\partial v^2} + 2u_x v_x \frac{\partial^2 \phi}{\partial u \partial v}.\end{aligned}\quad (10.41)$$

Similarly, we have

$$\begin{aligned}\frac{\partial^2 \phi}{\partial y^2} &= (u_y)^2 \frac{\partial^2 \phi}{\partial u^2} + (v_y)^2 \frac{\partial^2 \phi}{\partial v^2} + 2u_y v_y \frac{\partial^2 \phi}{\partial u \partial v} \\ &= (v_x)^2 \frac{\partial^2 \phi}{\partial u^2} + (u_x)^2 \frac{\partial^2 \phi}{\partial v^2} - 2v_x u_x \frac{\partial^2 \phi}{\partial u \partial v},\end{aligned}\quad (10.42)$$

where we have used the Cauchy–Riemann relations:  $u_x = v_y$ ,  $u_y = -v_x$ . Adding up the sides of the second lines of (10.41) and (10.42), we obtain

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} = \left[ (u_x)^2 + (u_y)^2 \right] \left( \frac{\partial^2 \phi}{\partial u^2} + \frac{\partial^2 \phi}{\partial v^2} \right).$$

The quantity inside the square brackets is equal to  $|f'(z)|^2$ , which is nonzero for analytic functions  $f(z)$ . As a consequence, we conclude that

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} = 0 \quad \Longleftrightarrow \quad \frac{\partial^2 \phi}{\partial u^2} + \frac{\partial^2 \phi}{\partial v^2} = 0. \quad \clubsuit$$

## 10.4 Applications in Physics and Engineering

### 10.4.1 Electric Potential Field in a Complicated Geometry

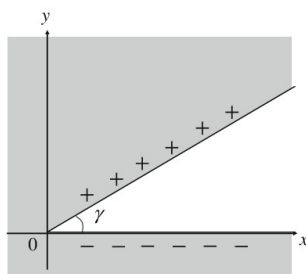
The Schwarz–Christoffel transformation is useful in mathematical physics, since it can be used to solve two-dimensional **Laplace equations** under certain boundary conditions. In fact, there are many physical systems that are described by Laplace’s equation subject to **Dirichlet** or **Neumann boundary conditions**. For example, Laplace’s equation can be used to describe heat conduction in a uniform medium, nonturbulent fluid flow, and an electrostatic field in a uniform system. In this subsection, we demonstrate how the Schwarz–Christoffel transformation works efficiently to solve such two-dimensional Laplace equations. It should be emphasized that our method is independent of the physical system being described. In the meantime, we apply the transformation to problems in electrostatics in order to illustrate

the method of solution, bearing in mind that these techniques are also applicable to problems involving other physical systems.

The general procedure for determining the electrostatic potential by using conformal mapping methods involves transforming a complicated charge-distribution geometry in the  $z$ -plane into a simple geometry in the  $w$ -plane. After solving the problem for the simpler geometry, the inverse transformation to the  $z$ -plane is applied to obtain the potential for the original geometry.

As a concrete example, we consider a metal block with a cut out wedge of angle  $\gamma$  as shown in Fig. 10.11. There is a vacuum inside the wedge. The block extends to  $\pm\infty$  in the direction perpendicular to the plane of the page. Since charge moves freely inside a metal, all of the charge placed in the conductor is distributed in such a way that the potential at all points along these edges is the same. We denote this potential by  $\phi_0$ , i.e., the system under consideration is subject to the Dirichlet boundary conditions given by

$$\phi(r, \theta = 0) = \phi(r, \theta = \gamma) = \phi_0. \quad (10.43)$$



**Fig. 10.11.** Wedge cut of a metal

Our objective is to evaluate the potential  $\phi(z)$  at points in the vacuum region inside the wedge (defined by  $0 \leq \arg(z) \leq \gamma$ ). This potential satisfies the Laplace equation, and thus, it can be determined by conformal mapping methods. For this purpose, we attempt to find the mapping that transforms the wedge in the  $z$ -plane onto the real axis of the  $w$ -plane. We know that the transformation of the real axis in the  $z$ -plane onto the wedge shown in Fig. 10.11 is given by the Schwarz–Christoffel transformation:

$$w = \beta + \alpha(z - x_1)^{-\theta_1/\pi}.$$

Therefore, the inverse mapping

$$z = x_1 + \left[ \frac{1}{\alpha}(w - \beta) \right]^{-\pi/\theta_1} \quad (10.44)$$

transforms the wedge in the  $w$ -plane with an internal angle  $-\theta_1$  onto the real axis of the  $z$ -plane. By interchanging  $z$  and  $w$  in (10.44), we obtain the mapping

$$w = u_1 + \left[ \frac{1}{\alpha}(z - \beta) \right]^{-\pi/\theta_1}, \quad (10.45)$$

which transforms the wedge in the  $z$ -plane onto the real axis of the  $w$ -plane. In order to apply this mapping to the configuration shown in Fig. 10.11, we set

$$\gamma = -\theta_1, \quad u_1 = \beta = 0, \quad \text{and} \quad \frac{1}{\alpha} = 1,$$

where  $\alpha$  is real. Then, the mapping in (10.45) becomes

$$w = z^{\pi/\gamma}. \quad (10.46)$$

**Remark.** It immediately follows that the mapping in (10.46) transforms the space within the wedge onto the upper half of the  $w$  plane. This is because points within the wedge that satisfy the condition  $0 < \arg(z) = \theta < \gamma$  are mapped onto  $w = r^{\pi/\gamma} e^{i\pi\theta/\gamma}$ , whose argument  $\pi\theta/\gamma$  takes values in the interval  $(0, \pi)$ .

Remember that the Dirichlet boundary condition is invariant under conformal mappings. Hence, the boundary condition of (10.43) is mapped to

$$\phi(u, v = 0) = \phi_0, \quad (10.47)$$

where  $v = 0$  is the image of the wedge. As noted earlier, the mapping in (10.46) transforms the problem of finding the potential in the region within the wedge in Fig. 10.11 to that of finding the potential in the upper half of the  $w$  plane due to a flat metal surface that extends along the entire  $u$ -axis and is maintained at a potential  $\phi_0$  by a uniformly distributed charge.

We now consider the “mapped” Laplace equation for the  $w$ -plane. Since all points on the surface of the flat plane are at the same potential, the potential all points  $(u, v)$  located at the same distance  $v$  above the plate is the same. Thus, the potential at any point must be independent of the value of  $u$  and the Laplace equation in the  $w$ -plane becomes

$$\frac{d^2\phi}{dv^2} = 0.$$

Integration of this differential equation followed by application of the boundary condition (10.47) yields

$$\phi(v) = \phi_0 + cv. \quad (10.48)$$

The constant  $c$  is obtained by using the property that the derivative of the potential (i.e., the electrostatic field) is a constant for a charged flat plate. Similar to  $\phi_0$ , the value of this constant field  $E_0$  depends on how much charge is distributed over a given area on the plate. With reference to (10.48),

$$\frac{\partial \phi}{\partial v} = c = -E_0, \quad \text{so that} \quad \phi(v) = \phi_0 - E_0 v.$$

In order to complete the analysis, the potential must be expressed in terms of the coordinates in the  $z$ -plane. From this expression

$$v = \text{Im}(w) = \text{Im}\left(r^{\pi/\gamma} e^{i\pi\theta/\gamma}\right) = r^{\pi/\gamma} \sin(\pi\theta/\gamma),$$

the potential is given by

$$\begin{aligned} \phi &= \phi_0 - E_0 r^{\pi/\gamma} \sin(\pi\theta/\gamma) \\ &= \phi_0 - E_0 (x^2 + y^2)^{\pi/(2\gamma)} \sin\left[\frac{\pi}{\gamma} \tan^{-1}\left(\frac{y}{x}\right)\right]. \end{aligned} \quad (10.49)$$

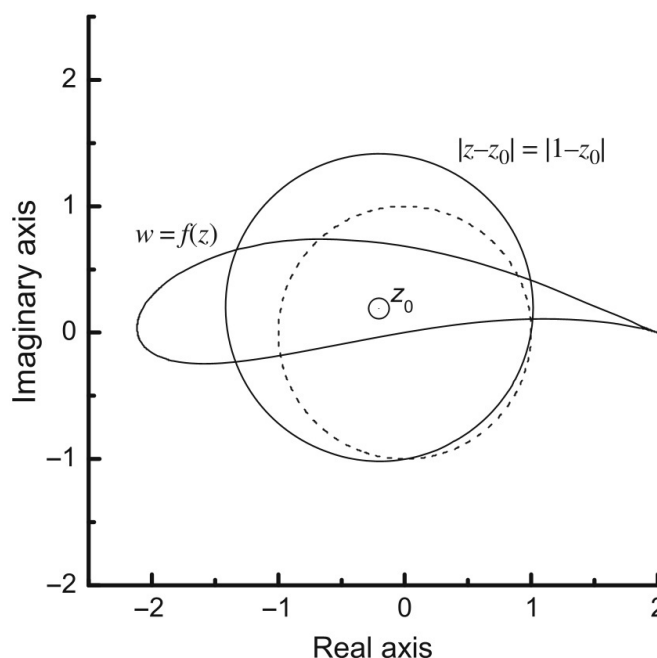
This is the final solution to the problem in question. We see from (10.49) that  $\phi = \phi(r, \theta)$  is constant when

$$r^{\pi/\gamma} \sin(\pi\theta/\gamma) = \text{const.}$$

This is the equation for a family of equipotential curves.

### 10.4.2 Joukowski Airfoil

Our final discussion related to the applications of conformal mappings concerns the **Joukowski transformation**, which is an important conformal



**Fig. 10.12.** The Joukowski transformation (10.50) of the circle  $|z - z_0| = |1 - z_0|$  with  $z_0 = (-0.2, 0.2)$  to the airfoil indicated by the *thick curve*

mapping that has been historically employed in the theory of **airfoil** design. Here, the term “airfoil” refers to the cross-sectional shape of a wing (or a propeller or a turbine). According to the literature on airfoil theory, any object with an angle of attack in a moving fluid generates a **lift**, a force perpendicular to the flow. Airfoils are designed as efficient shapes that increase the lift that the object generates. The Joukowski transformation maps a circle on the complex plane into a family of airfoil shapes called **Joukowski airfoils**, which simplify the analysis of two-dimensional fluid flows around an airfoil with a complicated geometry.

The Joukowski transformation  $w = f(z)$  is defined by

$$w = f(z) = z + \frac{1}{z}, \quad (10.50)$$

where  $z$  is located on a circle  $C$  that passes through the point  $z = 1$  and encloses the point  $z = -1$  as well as the origin  $z = 0$ . Note that the center of the circle, denoted by  $z_0$ , does not coincide with the origin, but is located close to the origin. In fact, the coordinates of  $z_0$  are variables, and changes in these variables alter the geometry of the resulting airfoil. An example of an airfoil generated by the transformation (10.50) is shown in Fig. 10.12, where  $z_0 = (-0.2, 0.2)$ . We see that the circle  $C : |z - z_0| = |1 - z_0|$  is mapped onto an airfoil indicated by a thick curve. The stream lines for a flow around the airfoil can be obtained by applying an inverse transformation to the streamlines for a flow around the circle and the latter can be easily evaluated.

## Part IV

---

### Fourier Analysis



## Fourier Series

---

**Abstract** A Fourier series is an expansion of a periodic function in terms of an infinite sum of sines and cosines. The use of a Fourier series allows us to break up an arbitrary periodic function into a set of simple terms that can be solved individually and then recombined in order to obtain the solution to the original problem with the desired level of accuracy. In this chapter, we place particular emphasis on the mean convergence property of a Fourier series (Sect. 11.2.1) and the conditions that are necessary for the series to be uniformly convergent (Sect. 11.3.1). Better understanding of convergence properties clarifies the reasons for the utility and the limit of validity of Fourier series expansion in mathematical physics.

### 11.1 Basic Properties

#### 11.1.1 Definition

**Fourier series** are infinite series consisting of trigonometric functions with a particular definition of expansion coefficients. They can be applied to almost all periodic functions whether the functions are continuous or not. With these expansion, physical phenomena involving some periodicity are reduced to a superposition of simple trigonometric functions, which helps us a great deal in arithmetic and practical aspects. In section We begin this with a description of the basic properties of Fourier series. We follow this by considering the convergence theory of Fourier series, which is the issue in the next section.

First of all, it is important to clarify the distinction between the following two concepts: trigonometric series and Fourier series.

♠ **Trigonometric series:**

The series

$$\frac{A_0}{2} + \sum_{n=1}^{\infty} (A_n \cos nx + B_n \sin nx)$$

is called a **trigonometric series**.

Here the set of coefficients  $\{A_n\}$  and  $\{B_n\}$  can be taken arbitrarily. (The expression  $A_0/2$  instead of  $A_0$  is just due to our convention.) Among the infinite choices of  $\{A_n\}$  and  $\{B_n\}$ , a specific definition of the coefficients noted below provides the Fourier series of a given function  $f(x)$ .

♠ **Fourier series:**

The series

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx) \quad (11.1)$$

is called a **Fourier series** of a function  $f(x)$  if and only if the coefficients are given by the **Euler–Fourier formula** expressed by

$$\begin{aligned} a_n &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos nx dx, \\ b_n &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin nx dx. \end{aligned} \quad (11.2)$$

Accordingly, a Fourier series is a specific kind of trigonometric series whose coefficients bear a definite relation (11.2) to some function  $f(x)$ . In (11.1) we have written the constant term as  $a_0/2$  rather than  $a_0$ , so that the expression for  $a_0$  is given by taking  $n = 0$  in (11.2). There is no  $b_0$  for  $\sin(0 \cdot x) = 0$ .

By definition, every Fourier series is a trigonometric series. However, the converse is not true, as demonstrated below.

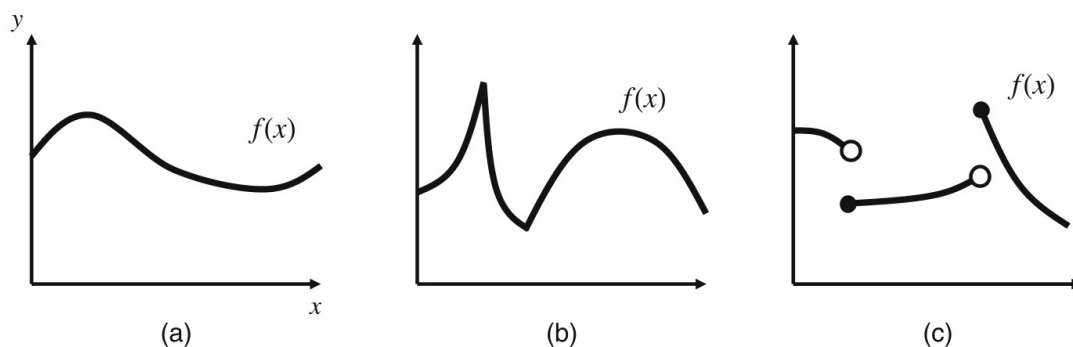
*Example* It is known that the trigonometric series given by

$$\sum_{n=2}^{\infty} \frac{\sin nx}{\log n}$$

is not a Fourier series. Indeed, no function can be related to the coefficient  $1/\log n$  via (11.2).

### 11.1.2 Dirichlet Theorem

Emphasis should be placed on the fact that the definition of Fourier series provides no information as to its convergence; thus the infinite series (11.1) may converge or diverge depending on the behavior of the function  $f(x)$ . This leads us to discuss which functions  $f(x)$  make the series (11.1) convergent. This issue is clarified in part by the following theorem (and by referring Fig. 11.1):



**Fig. 11.1.** (a) Continuous and smooth function. (b) Continuous but nonsmooth function. (c) Function with a finite number of discontinuities

♠ **Dirichlet theorem:**

If  $f(x)$  is periodic with the period  $2\pi$  and if  $f'(x)$  is continuous or at most have a finite number of discontinuity in  $[0, 2\pi]$ , then its Fourier series converges to

1.  $f(x)$ , if  $x$  is a point of continuity, or
2.  $\frac{f(x+0) + f(x-0)}{2}$ , if  $x$  is a point of discontinuity.

The set of conditions noted above is called **Dirichlet's conditions**. It is worthy to note that the Dirichlet conditions are *sufficient* but not *necessary*. That is, if the conditions are satisfied, the convergence of the series is guaranteed; but if they are not satisfied, the series may or may not converge. An exact proof of Dirichlet's theorem requires rather complicated calculations, which will be demonstrated in the next section.

*Remarks.*

1. The Dirichlet conditions do not require the continuity of  $f(x)$  within  $[0, 2\pi]$ .
2. Almost all periodic functions that we encounter in physical problems satisfy the Dirichlet conditions; therefore, the Fourier series expansion can be used almost regardless of its convergence.

It follows that if  $f(x)$  is continuous within  $[0, 2\pi]$  and satisfies Dirichlet's conditions, then the Fourier series of  $f(x)$  converges to  $f(x)$  at all the points within  $[0, 2\pi]$ . This means that the Fourier series of  $f(x)$  converges *uniformly* to  $f(x)$ . Once uniform convergence is ensured, we generally write

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx) \quad (11.3)$$

with the definition (11.2) for the coefficients. Consequently, if we form the Fourier series of  $f(x)$  without first examining its convergence to  $f(x)$ , we should write

$$f(x) \sim \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx) \quad (11.4)$$

instead of (11.3). The symbol “ $\sim$ ” in (11.4) means that the series on the right-hand side only *corresponds* to the function  $f(x)$  and can be replaced by the equality “ $=$ ” only if we succeed in proving that the infinite series converges uniformly to  $f(x)$ .

### 11.1.3 Fourier Series of Periodic Functions

Preceding arguments were limited to the case of periodic functions with period  $2\pi$ . But Fourier series expansions can apply to periodic functions whose periods differ from  $2\pi$ . This is seen by replacing  $x$  in (11.3) by  $(2\pi/\lambda)x$ , which transforms a series convergent in the interval  $[0, 2\pi]$  to another series convergent to  $[0, \lambda]$ . The resulting Fourier series is

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos nkx + b_n \sin nkx), \quad (11.5)$$

where  $k = 2\pi/\lambda$  and

$$a_n = \frac{2}{\lambda} \int_0^{\lambda} f(x) \cos nkx dx \quad \text{and} \quad b_n = \frac{2}{\lambda} \int_0^{\lambda} f(x) \sin nkx dx. \quad (11.6)$$

Obviously, these latter expressions can be reduced to the original definitions (11.1) and (11.2) by setting  $\lambda = 2\pi$ .

The expressions (11.5) and (11.6) become more concise by imposing the relations

$$\cos(nkx) = \frac{e^{inkx} + e^{-inkx}}{2}, \quad \sin(nkx) = \frac{e^{inkx} - e^{-inkx}}{2i}.$$

Then the Fourier series reads

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left( \frac{a_n - ib_n}{2} \right) e^{inkx} + \sum_{n=1}^{\infty} \left( \frac{a_n + ib_n}{2} \right) e^{-inkx}. \quad (11.7)$$

We rewrite the index  $n$  in the second sum by  $-n'$  to find

$$\begin{aligned} \sum_{n=1}^{\infty} \left( \frac{a_n + ib_n}{2} \right) e^{-inkx} &= \sum_{-n'=1}^{\infty} \left( \frac{a_{-n'} + ib_{-n'}}{2} \right) e^{in'kx} \\ &= \sum_{n'=-1}^{-\infty} \left( \frac{a_{n'} - ib_{n'}}{2} \right) e^{in'kx}, \end{aligned}$$

where the identities  $a_{-n} = a_n$  and  $b_{-n} = -b_n$  were used. As a result, we obtain a complex form of the Fourier series as

$$\begin{aligned} f(x) &\sim \frac{a_0}{2} + \sum_{n=1}^{\infty} \left( \frac{a_n - ib_n}{2} \right) e^{inkx} + \sum_{n=-1}^{-\infty} \left( \frac{a_n - ib_n}{2} \right) e^{inkx} \\ &= \sum_{n=-\infty}^{\infty} c_n e^{inkx}, \end{aligned} \quad (11.8)$$

with the definition

$$c_n = \frac{a_n - ib_n}{2}. \quad (11.9)$$

An explicit form of  $c_n$  is given by substituting the definition of  $a_n$  and  $b_n$ , given by (11.6), into (11.9) as

$$\begin{aligned} c_n &= \frac{1}{2} \left\{ \frac{2}{\lambda} \int_0^{\lambda} f(x) \cos(nkx) dx - \frac{2i}{\lambda} \int_0^{\lambda} f(x) \sin(nkx) dx \right\} \\ &= \frac{1}{\lambda} \int_0^{\lambda} f(x) e^{-inkx} dx. \end{aligned} \quad (11.10)$$

#### 11.1.4 Half-range Fourier Series

Fourier series expansions sometimes involve only sine or cosine terms. This actually occurs when the function being expanded is either even [ $f(-x) = f(x)$ ] or odd [ $f(-x) = -f(x)$ ] over the interval  $[-\lambda/2, \lambda/2]$ . When a given function is even or odd, unnecessary work in determining Fourier coefficients can be avoided. For instance, for an odd function  $f_o(x)$ , we have

$$\begin{aligned} a_n &= \frac{2}{\lambda} \int_{-\lambda/2}^{\lambda/2} f_o(x) \cos(nkx) dx \\ &= \frac{2}{\lambda} \left\{ \int_{-\lambda/2}^0 f_o(x) \cos(nkx) dx + \int_0^{\lambda/2} f_o(x) \cos(nkx) dx \right\} \\ &= \frac{2}{\lambda} \left\{ - \int_0^{\lambda/2} f_o(x) \cos(nkx) dx + \int_0^{\lambda/2} f_o(x) \cos(nkx) dx \right\} \\ &= 0 \quad (n = 0, 1, 2, \dots) \end{aligned} \quad (11.11)$$

and

$$\begin{aligned} b_n &= \frac{2}{\lambda} \left\{ \int_{-\lambda/2}^0 f_o(x) \sin(nkx) dx + \int_0^{\lambda/2} f_o(x) \sin(nkx) dx \right\} \\ &= \frac{4}{\lambda} \int_0^{\lambda/2} f_o(x) \sin(nkx) dx \quad (n = 0, 1, 2, \dots). \end{aligned} \quad (11.12)$$

Here we used the identities  $\cos(-nkx) = \cos(nkx)$  and  $\sin(-nkx) = -\sin(nkx)$ . Accordingly, we have

$$f_o(x) \sim \sum_{n=1}^{\infty} b_n \sin(nkx),$$

which is called the **Fourier sine series**.

Similarly, in the Fourier series corresponding to an even function  $f_e(x)$ , the same process yields

$$a_n = \frac{4}{\lambda} \int_0^{\lambda/2} f_e(x) \cos(nkx) dx \quad (n = 0, 1, 2, \dots) \quad (11.13)$$

and  $b_n = 0$  for all  $n$ . Accordingly, the Fourier series becomes

$$f_e(x) \sim \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(nkx),$$

which is called the **Fourier cosine series**.

Note that  $a_n$  and  $b_n$  given in (11.12) and (11.13) are computed in the interval  $[0, \lambda/2]$ , whose width is half of the period  $\lambda$ . Thus, the Fourier sine or cosine series of an odd or even function, respectively, is often called a **half-range Fourier series**. As discussed later, half-range Fourier series expansion is important from a practical viewpoint because it enables us to expand a nonperiodic function within its domain.

♠ **Theorem:**

If  $f(x)$  is an even or odd function and it is periodic with period  $\lambda$ , then the Fourier coefficients  $a_n$  and  $b_n$  become

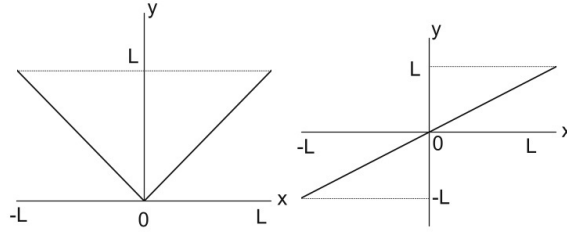
$$a_n = \frac{4}{\lambda} \int_0^{\lambda/2} f(x) \cos(nkx) dx, \quad b_n = 0 \quad \text{if } f(x) \text{ is even}$$

and

$$a_n = 0, \quad b_n = \frac{4}{\lambda} \int_0^{\lambda/2} f(x) \sin(nkx) dx \quad \text{if } f(x) \text{ is odd.}$$

### 11.1.5 Fourier Series of Nonperiodic Functions

A problem that arises quite often in applications is how to apply a Fourier series expansion to a function  $f(x)$  that is defined only on the interval  $[0, L]$ . In this case, nothing is said about the periodicity of  $f(x)$ . However, this does not prevent us from writing the Fourier series of  $f(x)$ , since the Euler–Fourier formulas (11.2) involve only the finite interval.



**Fig. 11.2.** Functions  $f_e(x)$  and  $f_o(x)$  defined in (11.14) and (11.15), respectively

As an example, we try to expand the function

$$f(x) = x \text{ for } [0, L]$$

as a Fourier series. In this case,  $f(x)$  is not periodic, but we can make it a periodic function by extending it as an even or odd function over  $[-L, L]$  and periodic with period  $2L$ . The respective definitions of  $f_e(x)$  and  $f_o(x)$  in  $[-L, L]$  are

$$f_e(x) = \begin{cases} -x & \text{for } -L \leq x < 0, \\ x & \text{for } 0 \leq x \leq L \end{cases} \quad (11.14)$$

and

$$f_o(x) = x \text{ for } -L \leq x \leq L, \quad (11.15)$$

whose profiles are shown in Fig. 11.2.

First, we consider the case of the even function  $f_e(x)$ . In terms of the Fourier cosine expansion, the coefficients  $a_0$  and  $a_n$  are given by

$$a_n = \frac{2}{L} \int_0^L f_e(x) \cos(nkx) dx = \frac{2L}{\pi^2} \frac{[(-1)^n - 1]}{n^2} = \begin{cases} \frac{-4L}{n^2\pi^2}, & n = 1, 3, \dots, \\ 0, & n = 2, 4, \dots \end{cases},$$

$$a_0 = \frac{2}{L} \int_0^L x dx = L.$$

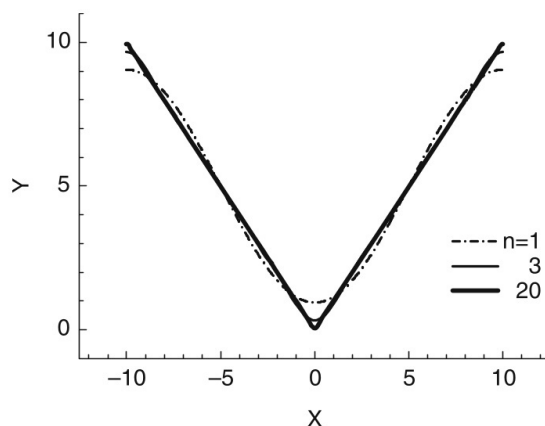
Here we have used  $kL = \pi$ . Hence, the cosine series becomes

$$f(x) = \frac{L}{2} - \frac{4L}{\pi^2} \sum_{n=1}^{\infty} \frac{1}{(2n-1)^2} \cos \frac{(2n-1)\pi x}{L}. \quad (11.16)$$

The partial sums of the series given in (11.16) are illustrated in Fig. 11.3. Although the original function  $f(x)$  is defined only within the interval  $[0, L]$ , the resulting Fourier series produces not only  $f(x)$  in  $[0, L]$ , but also the even extension  $f_e(x)$  with  $f_e(x) = f_e(x + 2L)$ .

Second, we look at the sine series of  $f_o$  given by (11.15). In this case,

$$b_n = \frac{2}{L} \int_0^L f_o(x) \sin(nkx) dx = \frac{2L}{\pi} \frac{(-1)^{n+1}}{n}$$

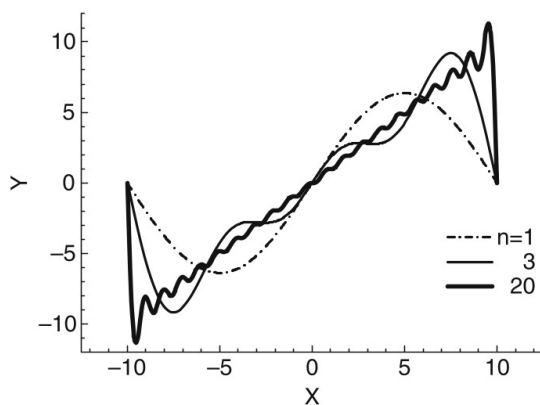


**Fig. 11.3.** A partial sum on the right-hand side of (11.16)

and the sine series is

$$f(x) = \frac{2L}{\pi} \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} \sin(nkx). \quad (11.17)$$

Figure 11.4 shows some partial sums of (11.17). As in the case of even extension, the Fourier series produces the odd extension  $f_o(x)$  with  $f_o(x) = f_e(x + 2L)$ .



**Fig. 11.4.** A partial sum on the right-hand side of (11.17)

### 11.1.6 The Rate of Convergence

We have had two kinds of Fourier series representations for  $f(x) = x$  in the interval  $[0, L]$ . This poses the following question: Does it make any difference which kind of Fourier series, (11.16) or (11.17), we use to represent  $f(x) = x$  in the interval  $[0, \lambda/2]$ ? Yes, it does. In the above-mentioned case, the even extension  $f_e(x)$  is more suitable than the odd extension  $f_o(x)$  for following two reasons.

The first reason concerns the rate of convergence of the resulting Fourier series. The coefficients given in (11.16) go as  $1/(2n-1)^2$ , whereas those in (11.17) go as  $1/n$ . Thus, the former series converges more quickly than the latter. The difference in the rate of convergence is due to the fact that the periodic extension of  $f_e(x)$  is continuous, but that of  $f_o(x)$  has discontinuities at odd multiples of  $L$ . In general, the Fourier coefficients of discontinuous functions decay as  $1/n$ , whereas those of continuous functions decay at least as rapidly as  $1/n^2$ . These observations as to the rate of convergence of the coefficient with respect to  $n$  can be formulated as follows:

♠ **Theorem:**

If  $f(x)$  and its first  $k$  derivatives satisfy the Dirichlet conditions on the interval  $[0, \lambda]$  and if the periodic extensions of  $f(x), f'(x), \dots, f^{(k-1)}(x)$  are all continuous, then the Fourier coefficients of  $f(x)$  decay at least as rapidly as  $1/n^{k+1}$ .

The second reason is that the Fourier series representation corresponding to the odd extension  $f_o(x)$  exhibits a small discrepancy from the original function  $f(x)$  around points of discontinuity of  $f_o(x)$ . This discrepancy is a **Gibbs phenomenon**, illustrated in Sect. 11.3.5. When an extension generates points of discontinuities, a **Gibbs phenomenon** will inevitably occur, which makes the resulting Fourier series representation highly unreliable in the vicinity of the discontinuity. Consequently, when performing half-range expansions of nonperiodic functions, the way of extension that renders the resulting function continuous (and smooth) over its domain is preferred.

### 11.1.7 Fourier Series in Higher Dimensions

It is important to generalize the Fourier series to more than one dimension. This generalization is especially useful in crystallography and solid-state physics, which deal with the three-dimensional periodic structures of atoms and molecules. To generalize to  $N$  dimensions, we first consider a special case in which an  $N$ -dimensional periodic function is a product of  $N$  one-dimensional periodic functions. That is, we take the  $N$  functions  $f^{(j)}(x)$  [ $j = 1, 2, \dots, N$ ] with period  $L_j$ :

$$f^{(j)}(x) = \sum_{n=-\infty}^{\infty} c_n^{(j)} e^{2\pi i n x / L_j}, \quad j = 1, 2, \dots, N, .$$

Let us define  $F(\mathbf{r})$  by the product of all the  $N$  functions  $f^{(j)}(x_j)$

$$\begin{aligned}
 F(\mathbf{r}) &= f^{(1)}(x_1)f^{(2)}(x_2)\cdots f^{(N)}(x_N) \\
 &= \sum_{n_1} \sum_{n_2} \cdots \sum_{n_N} c_{n_1}^{(1)} c_{n_2}^{(2)} \cdots c_{n_N}^{(N)} \cdot e^{2\pi i(n_1 x_1/L_1 + \cdots + n_N x_N/L_N)} \\
 &= \sum_{\mathbf{k}} C_{\mathbf{k}} e^{i\mathbf{k} \cdot \mathbf{r}}, \tag{11.18}
 \end{aligned}$$

where we have used the following new notation:

$$\begin{aligned}
 C_{\mathbf{k}} &= c_{n_1}^{(1)} c_{n_2}^{(2)} \cdots c_{n_N}^{(N)}, \\
 \mathbf{k} &= 2\pi(n_1/L_1, n_2/L_2, \cdots, n_N/L_N), \\
 \mathbf{r} &= (x_1, x_2, \cdots, x_N).
 \end{aligned}$$

We take (11.18) as the definition of the Fourier series for any periodic function of  $N$  variables. The definition of the coefficient  $C_{\mathbf{k}}$  can be developed for a general periodic function  $F(\mathbf{r})$  of  $N$  variables:

$$F(\mathbf{r}) = \sum_{\mathbf{k}} C_{\mathbf{k}} e^{i\mathbf{k} \cdot \mathbf{r}} \iff C_{\mathbf{k}} = \frac{1}{V} \int_V F(\mathbf{r}) e^{-i\mathbf{k} \cdot \mathbf{r}} d^N x, \tag{11.19}$$

where  $V \equiv L_1 L_2 \cdots L_N$  determines the smallest region of periodicity in  $N$  dimensions. When  $N = 1$  (11.19) obviously reduces to the Fourier series in one dimension.

**Remark.** The application of (11.19) requires some clarification regarding the region  $V$  of the integral. In one dimension, the shape of the smallest region of periodicity is unique, being simply a line segment of length  $L$ . In two or more dimensions, however, such regions can have a variety of shapes. For instance, in two dimensions, they can be rectangles, pentagons, hexagons, and so forth. Thus, we let  $V$  in (11.18) stand for a primitive cell of the  $N$ -dimensional lattice. This cell in three dimensions, which is important in solid-state physics, is called the **Wigner–Seitz cell**.

Recall that  $F(\mathbf{r})$  is a periodic function of  $\mathbf{r}$ . This means that when  $\mathbf{r}$  is changed by  $\mathbf{R}$ , where  $\mathbf{R}$  is a vector describing the boundaries of a cell, then we should get the same function:  $F(\mathbf{r} + \mathbf{R}) = F(\mathbf{r})$ . This implies that the periodicity of

$F(\mathbf{r})$  requires the vector  $\mathbf{k}$  to take only restricted directions and magnitudes. In fact, when replacing  $\mathbf{r}$  in (11.19) by  $\mathbf{r} + \mathbf{R}$ , we have

$$F(\mathbf{r} + \mathbf{R}) = \sum_{\mathbf{k}} C_{\mathbf{k}} e^{i\mathbf{k} \cdot (\mathbf{r} + \mathbf{R})} = \sum_{\mathbf{k}} (e^{i\mathbf{k} \cdot \mathbf{R}} \cdot C_{\mathbf{k}} e^{i\mathbf{k} \cdot \mathbf{r}}),$$

which is equal to  $F(\mathbf{r})$  if

$$e^{i\mathbf{k} \cdot \mathbf{R}} = 1, \quad \text{i.e.,} \quad \mathbf{k} \cdot \mathbf{R} = 2\pi \times (\text{integer}). \quad (11.20)$$

Equation (11.20) is a key relation in determining the allowed directions and magnitudes of the vector  $\mathbf{k}$ . In one-dimensional cases, the inner product reduces to  $\mathbf{k} \cdot \mathbf{R} = (2\pi n/L) \cdot L = 2\pi n$ ; thus (11.20) obviously holds true. In three dimensions, the vector  $\mathbf{R}$  is represented as  $\mathbf{R} = m_1 \mathbf{a}_1 + m_2 \mathbf{a}_2 + m_3 \mathbf{a}_3$ , where  $m_1$ ,  $m_2$ , and  $m_3$  are integers and  $\mathbf{a}_1$ ,  $\mathbf{a}_2$ , and  $\mathbf{a}_3$  are crystal axes, which are not generally orthogonal. Hence, condition (11.20) is satisfied when  $\mathbf{k} = n_1 \mathbf{b}_1 + n_2 \mathbf{b}_2 + n_3 \mathbf{b}_3$ , where  $n_1$ ,  $n_2$ , and  $n_3$  are integers and  $\mathbf{b}_1$ ,  $\mathbf{b}_2$ , and  $\mathbf{b}_3$  are the reciprocal lattice vectors defined by

$$\mathbf{b}_1 = \frac{2\pi(\mathbf{a}_2 \times \mathbf{a}_3)}{\mathbf{a}_1 \cdot (\mathbf{a}_2 \times \mathbf{a}_3)}, \quad \mathbf{b}_2 = \frac{2\pi(\mathbf{a}_3 \times \mathbf{a}_1)}{\mathbf{a}_1 \cdot (\mathbf{a}_2 \times \mathbf{a}_3)}, \quad \mathbf{b}_3 = \frac{2\pi(\mathbf{a}_1 \times \mathbf{a}_2)}{\mathbf{a}_1 \cdot (\mathbf{a}_2 \times \mathbf{a}_3)}.$$

In fact,

$$\mathbf{k} \cdot \mathbf{R} = \left( \sum_{i=1}^3 n_i \mathbf{b}_i \right) \cdot \left( \sum_{j=1}^3 m_j \mathbf{a}_j \right) = \sum_{i,j} n_i m_j \mathbf{b}_i \cdot \mathbf{a}_j,$$

and the reader may verify that  $\mathbf{b}_i \cdot \mathbf{a}_j = 2\pi \delta_{ij}$ . Thus we obtain

$$\mathbf{k} \cdot \mathbf{R} = 2\pi \sum_{j=1}^3 m_j n_j = 2\pi \times (\text{integer}).$$

## Exercises

1. Expand the following functions in Fourier series:

(i)  $f(x) = \sin ax$  on  $[-\pi, \pi]$ , where  $a$  is *not* an integer.

(ii)  $f(x) = \sin ax$  on  $[0, \pi]$ , where  $a$  is *not* an integer.

**Solution:** It is straightforward to obtain the results:

$$(i) \quad \sin ax = \frac{2}{\pi} \sin a\pi \sum_{n=1}^{\infty} (-1)^n \frac{n \sin nx}{a^2 - n^2}.$$

$$\begin{aligned}
 \text{(ii)} \quad \sin ax &= \frac{1 - \cos a\pi}{\pi} \left[ \frac{1}{a} + 2a \sum_{n=1}^{\infty} \frac{\cos 2nx}{a^2 - 4n^2} \right] \\
 &+ 2a \frac{1 + \cos a\pi}{\pi} \sum_{n=0}^{\infty} \frac{\cos(2n+1)x}{a^2 - (2n+1)^2}. \quad \clubsuit
 \end{aligned}$$

2. Expand the functions  $f(x) = \cos x$  on  $[0, \pi]$  in a Fourier sine series.

$$\text{Solution:} \quad \cos x = \frac{8}{\pi} \sum_{n=1}^{\infty} \frac{n \sin 2nx}{4n^2 - 1}. \quad \clubsuit$$

3. (i) Find the Fourier series of  $f(x) = x$  on the interval  $[-\pi, \pi]$ .

(ii) Prove that the identity  $\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots$ .

**Solution:**

$$\text{(i)} \quad f(x) = \sum_{n=1}^{\infty} \frac{2(-1)^{n+1}}{n} \sin nx.$$

(ii) If we substitute  $x = \pi/2$  in the series, we obtain

$$\frac{\pi}{2} = \sum_{n=1}^{\infty} \frac{2(-1)^{n+1}}{n} \sin \frac{n\pi}{2} = 2 \left( 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots \right),$$

which obviously gives the desired result.  $\clubsuit$

4. Expand the function  $f(x) = x^2$  into the Fourier cosine series on the domain  $[-\pi, \pi]$  and then prove that  $\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$  and  $\sum_{n=1}^{\infty} \frac{4(-1)^n}{n^2} = -\frac{\pi^2}{3}$ .

$$\text{Solution:} \quad \text{Straightforward calculations yield } x^2 = \sum_{n=1}^{\infty} \frac{4(-1)^n}{n^2} \cos nx.$$

By substituting  $x = \pi$  and  $x = 0$ , we obtain the desired equations.  $\clubsuit$

5. Determine both the cosine and sine series of  $f(x) = x^3 - x$  defined on the interval  $[0, 1]$ . Which series do you suppose converges more quickly?

**Solution:** We may set the even and odd extensions of  $f(x)$  over  $[-1, 1]$ , respectively, as

$$f_o(x) = x^3 - x \quad \text{for } -1 \leq x \leq 1,$$

and

$$f_e(x) = \begin{cases} -x^3 + x & \text{for } -1 \leq x < 0, \\ x^3 - x & \text{for } 0 \leq x \leq 1. \end{cases}$$

It follows that  $f_o(x)$  is smoother than  $f_e(x)$ ; namely,  $f_e(x)$  has a discontinuity in its derivative at  $\pm n$ . This implies that the sine series converges more rapidly than the cosine series. In fact, straightforward calculations yield the sine series

$$f_o(x) = \frac{12}{\pi^3} \sum_{n=1}^{\infty} \frac{(-1)^n}{n^3} \sin(n\pi x)$$

and the cosine series

$$f_e(x) = -\frac{1}{4} + \frac{2}{\pi^2} \sum_{n=1}^{\infty} \left\{ \frac{1 + (-1)^n 2}{n^2} + \frac{6}{n^4 \pi^2} [1 - (-1)^n] \right\} \cos(n\pi x).$$

The continuity of  $f_o(x)$  and its first derivatives leads to Fourier coefficients that decay as  $1/n^3$ , whereas the continuity of  $f_e(x)$  coupled with the discontinuity in  $f'_e(x)$  leads to Fourier coefficients that decay as  $1/n^2$ . ♣

## 11.2 Mean Convergence of Fourier Series

### 11.2.1 Mean Convergence Property

We know that Fourier series are endowed with a specific class of convergence called **mean convergence** (or **convergence in the mean**). This converging behavior is expressed by an integral:

$$\lim_{N \rightarrow \infty} \int_0^\lambda \left| f(x) - \sum_{n=-N}^N c_n e^{inkx} \right|^2 dx = 0. \quad (11.21)$$

Equation (11.21) applies regardless of the continuity and smoothness of the function  $f(x)$ , as far as  $f(x)$  is square-integrable.

*Remark.* From the viewpoint of **Hilbert space theory**, the relation (11.21) comes from the completeness property of the set of functions  $\{e^{inkx}\}$  in the sense of the **norm in the  $L^2$  space**. The  $L^2$  space is a specific kind of Hilbert space that is composed of a set of square-integrable functions  $f(x)$  expressed by

$$\int_a^b |f(x)|^2 dx < \infty.$$

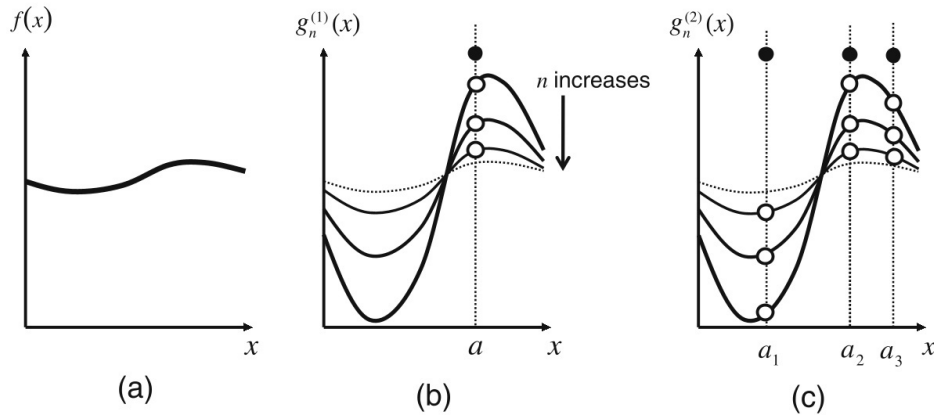
The **inner product**  $(f, g)$  and the norm  $\|f\|$  of elements  $f, g \in L^2$ , respectively, are given by

$$(f, g) = \int_a^b f(x)^* g(x) dx \quad \text{and} \quad \|f\| = (f, f) = \int_a^b |f(x)|^2 dx.$$

The mean convergence of the Fourier series [i.e., the equality in (11.21)] holds even when the integrand in (11.21) has a nonzero value at discrete points of  $x$ . This comes from the fact that the definition of the mean convergence is determined through integration, and that a finite number of discontinuities of the integrand do not contribute to the result of its integration. This is explained schematically in Fig. 11.5, in which we find

- $f(x)$  : a continuous function,
- $g_n^{(1)}(x)$  : a series that converges uniformly to  $f(x)$   
except at a point of discontinuity  $x = a$ .
- $g_n^{(2)}(x)$  : a series that converges uniformly to  $f(x)$   
except at points of discontinuity  $x = a_1, a_2, a_3, \dots$ .

As shown in Fig. 11.5, these three functions are distinct from one another. However, if we integrate the squared deviation between two of them followed by taking the limit  $n \rightarrow \infty$ , we have



**Fig. 11.5.** Sketches of a continuous function  $f(x)$ , a series of functions  $g_n^{(1)}(x)$  converging to  $f(x)$  except at a discontinuity, and a similar series of functions  $g_n^{(2)}(x)$  having several discontinuities. Series  $\{g_n^{(1)}(x)\}$  and  $\{g_n^{(2)}(x)\}$  both converge *in the mean* to  $f(x)$

$$\lim_{n \rightarrow \infty} \int_0^\lambda |f(x) - g_n^{(1)}(x)|^2 dx = \lim_{n \rightarrow \infty} \int_0^\lambda |f(x) - g_n^{(2)}(x)|^2 dx = 0. \quad (11.22)$$

This is because the area surrounded by two of them vanishes with  $n \rightarrow \infty$ , i.e., the area right below (or above) points of discontinuity are zero owing to their discreteness. Thus, the series  $g_n^{(1)}(x)$  and  $g_n^{(2)}(x)$  both converge to  $f(x)$  in the mean regardless of their discrepancy from  $f(x)$  at points of discontinuity.

### 11.2.2 Dirichlet and Fejér Integrals

It is pedagogical to give an alternative exposition of mean convergence of Fourier series, which is based on the two important concepts: **Dirichlet's integral** and **Fejér's integral**.

Consider the partial sum  $S_N(x)$  of the Fourier series of  $f(x)$  expressed by

$$S_N(x) = \sum_{n=-N}^N c_n e^{inkx}$$

and its arithmetic mean

$$\sigma_N(x) = \frac{1}{N+1} (S_0 + S_1 + \cdots + S_N). \quad (11.23)$$

After some algebra, we obtain their integral representations as given below (see Exercises 1 and 2 in 11.2.2 for references).

#### ♠ Dirichlet integral:

$$S_N(x) = \frac{1}{\lambda} \int_{-x}^{\lambda-x} f(t+x) \frac{\cos(Nkt) - \cos\{(N+1)kt\}}{1 - \cos(kt)} dt. \quad (11.24)$$

#### ♠ Fejér integral:

$$\sigma_N(x) = \frac{1}{(N+1)\lambda} \int_{-\lambda/2}^{\lambda/2} f(t+x) \frac{\sin^2 \frac{N+1}{2} kt}{\sin^2 \frac{1}{2} kt} dt. \quad (11.25)$$

**Remark.** Note the distinctive difference between the convergence of  $S_N$  and that of  $\sigma_N$ . Whereas  $\lim_{N \rightarrow \infty} S_N = S$  implies  $\sigma_N \rightarrow S$ , the converse does not generally hold true; in fact,  $\sigma_N$  may converge even when  $S_N$  diverges. A typical example is the case of the numerical sequence  $u_n = (-1)^n$ , where,  $S_N = \sum u_n$  does not converge because  $S_{2N} = 0$  and  $S_{2N+1} = 1$ , whereas  $\sigma_N = \sum S_n / (N+1)$  converges to  $1/2$ .

By putting  $f(x) \equiv 1$  in (11.25), we have the following notation:

♠ **Dirichlet kernel:** The function

$$D_N(t) \equiv \frac{1}{N+1} \cdot \frac{\sin^2 \frac{N+1}{2} kt}{\sin^2 \frac{1}{2} kt} \quad (11.26)$$

is called the **Dirichlet kernel**, which satisfies the identity:

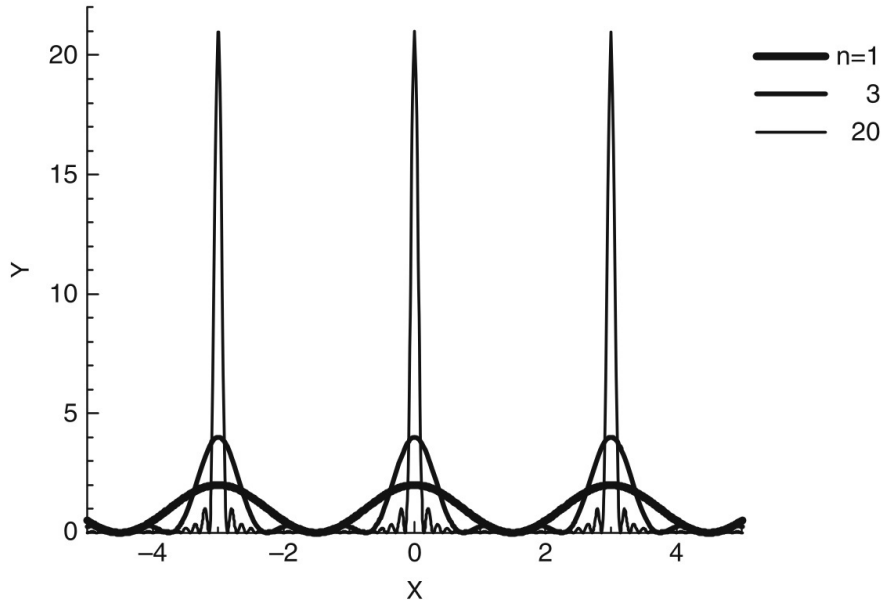
$$1 = \frac{1}{\lambda} \int_{-\lambda/2}^{\lambda/2} D_N(t) dt. \quad (11.27)$$

The derivation of the identity (11.27) is straightforward. When  $f(x) \equiv 1$ , we have  $f(t+x) = 1$ ,  $c_0 = 1$ , and  $c_n = 0$  ( $|n| \geq 1$ ), which obviously yield  $S_N = 1$  for arbitrary  $N$  and thus  $\sigma_N = 1$ . Substitute this into (11.25) to obtain the identity (11.27). Figure 11.6 plots the behavior of  $D_N(t)$  with increasing  $N$ ; it shows maxima at  $t = 0, \pm\lambda, \pm2\lambda, \dots$ , and the magnitude of the maxima become singular when  $N \rightarrow \infty$ .

From (11.25) and (11.27), we arrive at the key relation

$$\sigma_N(x) - f(x) = \frac{1}{\lambda} \int_{-\lambda/2}^{\lambda/2} \{f(t+x) - f(x)\} D_N(t) dt. \quad (11.28)$$

If  $f(x)$  is continuous (piecewise, at least), the integral in (11.28) can be made arbitrarily small by taking a sufficiently large  $N$  (see Exercise 3 below). To be precise, there exists an  $m$  for each  $\varepsilon > 0$  such that



**Fig. 11.6.** Dirichlet's kernel  $D_N(t)$  defined in (11.26)

$$N > m \Rightarrow |\sigma_N(x) - f(x)| < \varepsilon. \quad (11.29)$$

This clearly means that  $\sigma_N(x)$  converges uniformly to  $f(x)$  if  $f(x)$  is continuous.

As is shown later, the result (11.29) immediately yields the mean convergence of the Fourier series to  $f(x)$ .

### 11.2.3 Proof of the Mean Convergence of Fourier Series

We are now in a position to prove the mean convergence property of Fourier series.

The function  $\sigma_N(x)$  can be expressed as a trigonometric polynomial, since it consists of  $N$ 's trigonometric polynomials  $S_0, S_1, \dots, S_N$  as given by (11.23). Hence, (11.29) implies the existence of a trigonometric series that converges uniformly to  $f(x)$ . [This is simply **Fejér's theorem** (see Sect. 11.3.2).] Thus we have

$$\sigma_N(x) = \sum_{n=-N}^N \gamma_n e^{inkx},$$

where all the coefficients  $\{\gamma_n\}$  have to be determined.

We now make use of the fact that for any choice of  $\{\gamma_n\}$ , the inequality

$$\int_0^\lambda \left| f(x) - \sum_{n=-N}^N \gamma_n e^{inkx} \right|^2 dx \geq \int_0^\lambda \left| f(x) - \sum_{n=-N}^N c_n e^{inkx} \right|^2 dx$$

holds true with the Fourier coefficients  $\{c_n\}$  of  $f(x)$ . (See the discussion in Sect. 11.2.4 for the proof.) Taking the limit  $N \rightarrow \infty$  yields

$$\lim_{N \rightarrow \infty} \int_0^\lambda |f(x) - \sigma_N(x)|^2 dx \geq \lim_{N \rightarrow \infty} \int_0^\lambda \left| f(x) - \sum_{n=-N}^N c_n e^{inkx} \right|^2 dx. \quad (11.30)$$

Let  $f(x)$  be continuous (piecewise, at least). Then the left-hand side vanishes owing to the uniform convergence of  $\sigma_N(x)$  to  $f(x)$  at continuous points  $x$  of  $f(x)$  (A finite number of discontinuous points of  $f(x)$  makes no contribution to the integral.) Eventually, we come to the desired conclusion:

$$\lim_{N \rightarrow \infty} \int_0^\lambda \left| f(x) - \sum_{n=-N}^N c_n e^{inkx} \right|^2 dx = 0,$$

which is a restatement of (11.21).

### 11.2.4 Parseval Identity

The mean convergence property of Fourier series can be represented by a more concise expression, called the **Parseval identity**. We first note the main conclusion of this subsection and then go on to its proof. For simplicity of notation, we use the following short form:

$$\frac{1}{\lambda} \int_0^\lambda f(x)g^*(x)dx \equiv (f, g).$$

♠ **Parseval identity:**

A necessary and sufficient condition for the mean convergence of the Fourier series of  $f(x)$  is given by

$$(f, f) = \sum_{n=-\infty}^{\infty} |c_n|^2,$$

which is called the **Parseval identity**.

To prove the above statement, we assume  $f(x)$  to be square-integrable and consider the total squared error of  $f(x)$  relative to the series of exponential functions:

$$\mathcal{E}_N = \frac{1}{\lambda} \int_0^\lambda \left| f(x) - \sum_{n=-N}^N \gamma_n e^{inkx} \right|^2 dx, \quad (11.31)$$

whose variables are  $N$  and the sequence  $\{\gamma_n\}$  consisting of complex numbers. Term-by-term integration of (11.31) yields

$$\begin{aligned} \mathcal{E}_N &= (f, f) - \sum_{n=-N}^N \gamma_n^* (f, e^{inkx}) - \sum_{n=-N}^N \gamma_n (f, e^{inkx})^* \\ &\quad + \sum_{m,n=-N}^N \gamma_m^* \gamma_n (e^{inkx}, e^{imkx}) \\ &= (f, f) - \sum_{n=-N}^N (\gamma_n^* c_n + \gamma_n c_n^*) + \sum_{n=-N}^N \gamma_n^* \gamma_n \\ &= (f, f) + \sum_{n=-N}^N |\gamma_n - c_n|^2 - \sum_{n=-N}^N |c_n|^2. \end{aligned} \quad (11.32)$$

Here we have used the orthonormality of imaginary exponentials,  $(e^{inkx}, e^{imkx}) = \delta_{m,n}$ , as well as the definitions of the Fourier coefficient  $c_n = (f, e^{inkx})$ . Note that  $(f, f)$  appearing in (11.32) is nonnegative because

$$(f, f) = \frac{1}{\lambda} \int_0^\lambda |f(x)|^2 dx \geq 0.$$

Hence,  $\mathcal{E}_N$  becomes minimal when  $\gamma_n = c_n$  and its minimum value reads

$$\min\{\mathcal{E}_N\} = (f, f) - \sum_{n=-N}^N |c_n|^2. \quad (11.33)$$

We are now ready to complete our proof. Recall that the mean convergence of the Fourier series for  $f(x)$  is defined by

$$\lim_{N \rightarrow \infty} \int_0^\lambda \left| f(x) - \sum_{n=-N}^N c_n e^{inkx} \right|^2 dx = 0. \quad (11.34)$$

From (11.31) and (11.33), we see that the definition of the mean convergence (11.34) is rewritten as

$$\lim_{N \rightarrow \infty} \min\{\mathcal{E}_N\} = 0, \quad (11.35)$$

or equivalently,

$$(f, f) = \sum_{n=-\infty}^{\infty} |c_n|^2. \quad (11.36)$$

Relation (11.36) is thus a necessary and sufficient condition for satisfying the mean convergence of the Fourier series to  $f(x)$ . Since Parseval's identity applies to any square-integrable function  $f$ , Fourier series for the functions  $f$  surely converge in the mean to  $f(x)$ .

### 11.2.5 Riemann–Lebesgue Theorem

As by-products of the argument in 11.2.4, we obtain the following two important properties regarding the Fourier series expansion. The first is the **Bessel inequality**

$$(f, f) \geq \sum_{n=-N}^N |c_n|^2. \quad (11.37)$$

This is obtained from the fact that  $\min\{\mathcal{E}_N\}$  given in (11.33) is nonnegative. Here we can let  $N \rightarrow \infty$  in (11.37), because the right-hand side of (11.37) forms a monotonically increasing sequence that is bounded by its left-hand side. Then we obtain

$$(f, f) \geq \sum_{n=-\infty}^{\infty} |c_n|^2. \quad (11.38)$$

We further note that the series on the right-hand side of (11.38) necessarily converges, since it is nondecreasing and bounded from above. Consequently, we arrive at the second important property to be noted:

$$\lim_{n \rightarrow \infty} c_n = 0. \quad (11.39)$$

Separating the real and imaginary parts in (11.39), we eventually find the second point to be noted:

♠ **Riemann–Lebesgue theorem:**

If  $f(x)$  is square-integrable on the interval  $[0, \lambda]$ , then

$$\lim_{n \rightarrow \infty} \int_0^\lambda f(x) \cos(nkx) dx = 0, \quad \lim_{n \rightarrow \infty} \int_0^\lambda f(x) \sin(nkx) dx = 0.$$

## Exercises

1. Derive the expressions (11.24) and (11.25) regarding the Dirichlet and Fejér integrals, respectively.

**Solution:** From the definition of  $c_n$ , the partial sum  $S_N(x)$  yields its integral form:

$$\begin{aligned} S_N(x) &= \sum_{n=-N}^N \left\{ \frac{1}{\lambda} \int_0^\lambda f(t) e^{-inkt} dt \right\} \cdot e^{inkx} \\ &= \frac{1}{\lambda} \int_0^\lambda f(t) \left\{ \sum_{n=-N}^N e^{-ink(t-x)} \right\} dt \\ &= \frac{1}{\lambda} \int_{-x}^{\lambda-x} f(t+x) \left( \sum_{n=-N}^N e^{-inkt} \right) dt. \end{aligned} \quad (11.40)$$

The finite series of exponential terms reads

$$\begin{aligned} \sum_{n=-N}^N e^{-inkt} &= e^{-iNkt} \sum_{n=0}^{2N} e^{inkt} = e^{-iNkt} \cdot \frac{1 - e^{i(2N+1)kt}}{1 - e^{ikt}} \\ &= \frac{\cos(Nkt) - \cos\{(N+1)kt\}}{1 - \cos(kt)}. \end{aligned}$$

Substituting this in (11.40) yields Dirichlet's integral (11.24). Moreover, its arithmetic mean reduces to Fejér's integral (11.25) as demonstrated by

$$\begin{aligned} \sigma_N(x) &= \frac{1}{N+1} \{S_0 + S_1 + \cdots + S_N\} \\ &= \frac{1}{(N+1)\lambda} \int_{-x}^{\lambda-x} f(t+x) \frac{1 - \cos\{(N+1)kt\}}{1 - \cos(kt)} dt \\ &= \frac{1}{(N+1)\lambda} \int_{-\lambda/2}^{\lambda/2} f(t+x) \frac{\sin^2 \frac{N+1}{2} kt}{\sin^2 \frac{1}{2} kt} dt. \end{aligned} \quad (11.41)$$

In the last line of (11.41), the interval of the integration from  $[-x, \lambda - x]$  to  $[-\lambda/2, \lambda/2]$  is replaced by taking account of the periodicity of the integrand. ♣

2. Prove that  $\sigma_N(x)$  uniformly converges to  $f(x)$  by postulating the continuity of  $f(x)$ .

**Solution:** Recall that the continuity of  $f(x)$  allows us to determine a  $\delta$  that satisfies

$$|x - x'| < \delta \Rightarrow |f(x) - f(x')| < \varepsilon \quad (11.42)$$

for an arbitrary small  $\varepsilon$  to be positive. Further, owing to its continuity, the function  $f(x)$  is bounded as  $|f(x)| < M$  with an appropriate constant  $M$ . We divide the range of integration given in (11.28) as  $\int_{-\lambda/2}^{\lambda/2} = \int_{-\lambda/2}^{-\delta} + \int_{-\delta}^{\delta} + \int_{\delta}^{\lambda/2}$  and use the inequality (11.42) to obtain the middle term:

$$\begin{aligned} \left| \int_{-\delta}^{\delta} \{f(t+x) - f(x)\} D_N(t) dt \right| &\leq \int_{-\delta}^{\delta} |f(t+x) - f(x)| D_N(t) dt \\ &\leq \varepsilon \int_{-\delta}^{\delta} D_N(t) dt \leq \varepsilon \lambda \end{aligned} \quad (11.43)$$

and

$$\begin{aligned} &\left| \left( \int_{-\lambda/2}^{-\delta} + \int_{\delta}^{\lambda/2} \right) \{f(t+x) - f(x)\} D_N(t) dt \right| \\ &\leq \left( \int_{-\lambda/2}^{-\delta} + \int_{\delta}^{\lambda/2} \right) \{|f(t+x)| + |f(x)|\} \frac{\sin^2 \frac{N+1}{2} kt}{(N+1) \sin^2 \frac{1}{2} kt} dt \\ &\leq \left( \int_{-\lambda/2}^{-\delta} + \int_{\delta}^{\lambda/2} \right) 2M \cdot \frac{dt}{(N+1) \sin^2(k\delta/2)} \\ &\leq \frac{2M\lambda}{(N+1) \sin^2(k\delta/2)}. \end{aligned} \quad (11.44)$$

From (11.28), (11.43), and (11.44), we obtain

$$|\sigma_N(x) - f(x)| < \varepsilon + \frac{2M}{(N+1) \sin^2(k\delta/2)}.$$

Taking the limit  $N \rightarrow \infty$  and fixing the small quantity  $\delta$ , the second term vanishes. We thus conclude that

$$\lim_{N \rightarrow \infty} |\sigma_N(x) - f(x)| = 0. \quad \clubsuit \quad (11.45)$$

## 11.3 Uniform Convergence of Fourier series

### 11.3.1 Criterion for Uniform and Pointwise Convergence

We know that the Fourier series of  $f(x)$  converges in the mean to  $f(x)$  as far as  $f(x)$  is square-integrable. However, the mean convergence of the Fourier series provides no information as to its **uniform convergence**. In order for the Fourier series to converge (uniformly or pointwise) to the original function  $f(x)$ , several conditions regarding continuity and periodicity of  $f(x)$  have to be satisfied. These are formally stated in the following two theorems:

♠ **Uniform convergence of Fourier series:**

The Fourier series of a continuous, piecewise smooth, and periodic function  $f(x)$  converges to  $f(x)$  absolutely and uniformly.

♠ **Pointwise convergence of Fourier series:**

The Fourier series of a piecewise smooth and periodic function  $f(x)$  (continuous or discontinuous) converges to:

- (i)  $f(x)$  at any point of continuity, and
- (ii)  $\frac{f(x+0) + f(x-0)}{2}$  at any point of discontinuity.

Our main concern in this section is to prove these two theorems, and we follow this by demonstrating several important features of Fourier series that occur at discontinuous points of  $f(x)$ .

*Remark.* Observe that the above theorems are consistent with the conclusion of the **Dirichlet theorem** given in Sect. 11.1.2; the latter says that a Fourier series representation becomes identical to  $f(x)$  provided that  $f(x)$  is periodic, continuous, and further smooth (piecewise, at least).

### 11.3.2 Fejér theorem

The theorems given in the previous subsection clearly exhibit *sufficient* conditions for the Fourier series to converge. It is pedagogical to compare them with the **Fejér theorem**:

♠ **Fejér theorem:**

Any continuous and periodic function  $f(x)$  with a period  $\lambda$  can be reproduced by an infinite trigonometric series

$$\lim_{N \rightarrow \infty} \left| f(x) - \sum_{n=-N}^N \gamma_n e^{inkx} \right| = 0 \quad \text{for all } x, \quad (11.46)$$

with an appropriate choice of the set of expansion coefficients  $\{\gamma_n\}$ .

At first glance, Fejér's theorem appears to ensure the uniform convergence of the Fourier series. However, this is not the case at all; the sequence of the optimal coefficients  $\{\gamma_n\}$  satisfying (11.46) cannot in general be replaced by the Fourier coefficients  $\{c_n\}$  defined by

$$c_n = \frac{1}{\lambda} \int_0^\lambda f(x) e^{-inkx} dx.$$

In fact, even when  $f(x)$  is continuous and periodic, its Fourier series may diverge at discrete points, as is expressed by

$$\lim_{N \rightarrow \infty} \left| f(x) - \sum_{n=-N}^N c_n e^{inkx} \right| = \infty \quad \text{at some points } x. \quad (11.47)$$

Hence, Fejér's theorem does not guarantee the uniform convergence of the Fourier series representation. Equation (11.47) also suggests that the continuity and periodicity of  $f(x)$  are only necessary but not sufficient conditions, for the uniform convergence of its Fourier series to the original function  $f(x)$ .

### 11.3.3 Proof of Uniform Convergence

We are now in a position to prove the criterion for uniform convergence of the Fourier series  $\sum_{n=-\infty}^{\infty} c_n e^{inx}$  to  $f(x)$ . The proof that is presented below is based on the mean convergence property of Fourier series. Recall that the mean convergence of Fourier series is expressed by

$$\int_{-\infty}^{\infty} \left| f(x) - \sum_{n=-\infty}^{\infty} c_n e^{inx} \right|^2 dx = 0. \quad (11.48)$$

In general, the relation

$$\lim_{N \rightarrow \infty} \int_a^b \left| \sum_{n=0}^N u_n(x) \right|^2 dx = 0$$

means

$$\lim_{N \rightarrow \infty} \left\{ \sum_{n=0}^N u_n(x) \right\} = 0$$

if and only if the infinite series  $\sum_{n=0}^{\infty} u_n(x)$  converges uniformly to a certain function of  $x$  within the range of integration  $[a, b]$ . Therefore, in order to obtain the desired equality

$$f(x) = \sum_{n=-\infty}^{\infty} c_n e^{inkx}$$

for any  $x \in [0, \lambda]$ , we must seek the condition that the infinite series  $\sum_{n=-\infty}^{\infty} c_n e^{inkx}$  converges uniformly to some function of  $x$  [not necessarily to  $f(x)$ ]. Thus, we rewrite the Fourier coefficient  $c_n$  as

$$\begin{aligned} c_n &= \frac{1}{\lambda} \int_0^{\lambda} f(x) e^{-inkx} dx \\ &= \frac{1}{-ink\lambda} [f(x) e^{-inkx}]_0^{\lambda} + \frac{1}{ink\lambda} \int_0^{\lambda} f'(x) e^{-inkx} dx \\ &= \frac{1}{ink} \cdot \frac{1}{\lambda} \int_0^{\lambda} f'(x) e^{-inkx} dx = \frac{c'_n}{ink}, \end{aligned} \quad (11.49)$$

where  $c'_n$  is the Fourier coefficient of the derivative  $f'(x)$ . Here  $f(x)$  is assumed to be periodic, e.g.,  $f(0) = f(\lambda)$  and  $k\lambda = 2\pi$ . We further assume that  $f(x)$  is continuous and smooth (piecewise, at least) on the interval  $[0, \lambda]$ . Then,  $f'(x)$  is continuous (or piecewise continuous) to yield Parseval's identity:

$$\int_0^{\lambda} |f'(x)|^2 dx = \sum_{n=-\infty}^{\infty} |c'_n|^2 \equiv A,$$

where  $A$  is a constant. Observe that

$$\sum_{n=-\infty}^{\infty} |c_n| = \sum_{n=-\infty}^{\infty} \left| \frac{c'_n}{ink} \right| = \sum_{n=-\infty}^{\infty} \frac{|c'_n|}{nk}. \quad (11.50)$$

From the Schwartz inequality, it follows that

$$\begin{aligned} \sum_{n=-\infty}^{\infty} \frac{|c'_n|}{nk} &= \sum_{n=-\infty}^{\infty} \sqrt{\frac{1}{n^2 k^2}} \sqrt{|c'_n|^2} \leq \sqrt{\sum_{n=-\infty}^{\infty} \frac{1}{n^2 k^2}} \sqrt{\sum_{n=-\infty}^{\infty} |c'_n|^2} \\ &= \frac{A}{k} \sqrt{\sum_{n=-\infty}^{\infty} \frac{1}{n^2}}. \end{aligned} \quad (11.51)$$

It follows that  $\sum_{n=1}^{\infty} (1/n^2)$  is convergent (See the remark below). Hence, from (11.50) and (11.51), we see that  $\sum_{n=-\infty}^{\infty} |c_n|$  converges. This implies that  $\sum_{n=-\infty}^{\infty} c_n e^{inkx}$  converges uniformly to a certain function on  $[0, \lambda]$  since  $|c_n e^{inkx}| \leq |c_n|$  for all  $n$  on  $[0, \lambda]$ . (See Sect. 3.3.1 for the criteria for uniform convergence.) This completes our proof.

*Remark.* That the series  $\sum_{n=1}^{\infty} (1/n^2)$  is convergent is verified as follows: set  $A_{2^{k+1}-1}$  to be a partial sum consisting of the first  $2^{k+1} - 1$  terms. Then we have

$$\begin{aligned} A_{2^{k+1}-1} &= 1 + \left( \frac{1}{2^2} + \frac{1}{3^2} \right) + \left( \frac{1}{4^2} + \cdots + \frac{1}{7^2} \right) + \cdots \\ &\quad + \left[ \frac{1}{(2^k)^2} + \cdots + \frac{1}{(2^{k+1}-1)^2} \right] \\ &< 1 + \frac{1}{2^2} \times 2 + \frac{1}{4^2} \times 4 + \cdots + \frac{1}{(2^k)^2} \times 2^k \\ &< \sum_{j=0}^k \left( \frac{1}{2} \right)^j = \frac{1 - (1/2)^{k+1}}{1 - (1/2)} < 2. \end{aligned}$$

This means that  $A_{2^{k+1}-1}$  for any  $k$  is bounded above. Furthermore, the sequence  $(A_m)$  is monotonically increasing. Hence,  $(A_m)$  converges in the limit of  $m \rightarrow \infty$ , which completes the proof.

#### 11.3.4 Pointwise Convergence at Discontinuous Points

This subsection gives an account of the second criterion in Sect. 11.3.1, which is restated below.

##### ♠ Pointwise convergence at discontinuities:

When a function  $f(x)$  is piecewise continuous and piecewise smooth, its Fourier series converges pointwise to  $\{f(x+0) - f(x-0)\}/2$  at a point of discontinuity.

This theorem can be proven in the following manner. It readily follows from (11.24) that the partial sum of the Fourier series  $S_N(x)$  is expressed by

$$S_N(x) = \frac{1}{2i\lambda} \int_{-x}^{\lambda-x} f(x+t) \frac{e^{i(N+\frac{1}{2})kt} - e^{-i(N+\frac{1}{2})kt}}{\sin(\frac{1}{2}kt)} dt.$$

We rewrite this as

$$\begin{aligned} S_N(x) &= \frac{1}{2i\lambda} \int_{-x}^{\lambda-x} f(x+t) \frac{e^{i\frac{1}{2}kt}}{\sin(\frac{1}{2}kt)} \cdot e^{iNkt} dt \\ &\quad - \frac{1}{2i\lambda} \int_{-x}^{\lambda-x} f(x+t) \frac{e^{-i\frac{1}{2}kt}}{\sin(\frac{1}{2}kt)} \cdot e^{-iNkt} dt \\ &= \frac{1}{2i\lambda} \int_{-x}^{\lambda-x} \{f(x+t) + f(x-t)\} \frac{e^{i\frac{1}{2}kt}}{\sin(\frac{1}{2}kt)} \cdot e^{iNkt} dt. \end{aligned}$$

Here we have set  $t \rightarrow -t$  in the second integral in the first line. Further,

$$S_N(x) = \frac{1}{2i\lambda} \int_{-x}^{\lambda-x} g(t) e^{iNkt} dt + \frac{f(x+0) + f(x-0)}{2i\lambda} \int_{-x}^{\lambda-x} \frac{e^{i\frac{1}{2}kt} e^{iNkt}}{\sin(\frac{1}{2}kt)} dt, \quad (11.52)$$

where we have introduced the notation

$$g(t) = \{f(x+t) - f(x+0) + f(x-t) - f(x-0)\} \frac{e^{i\frac{1}{2}kt}}{\sin(\frac{1}{2}kt)}.$$

The second term in (11.52) can be simplified via the relation

$$\frac{1}{i\lambda} \int_{-x}^{\lambda-x} \frac{e^{i\frac{1}{2}kt} e^{iNkt}}{\sin(\frac{1}{2}kt)} dt = 1.$$

(See Exercise 3 in Sect. 11.3 for its derivation.) Substituting this into (11.52), we get

$$S_N(x) = \frac{1}{2i\lambda} \int_{-x}^{\lambda-x} g(t) e^{iNkt} dt + \frac{f(x+0) + f(x-0)}{2}. \quad (11.53)$$

If the integration term in (11.53) vanishes with  $N \rightarrow \infty$ , we will successfully obtain the desired result. In fact, when  $g(t)$  is piecewise continuous in the interval  $[-x, \lambda - x]$ , the integral in (11.53) vanishes owing to the Riemann–Lebesgue theorem (see Sect. 11.2.5). The remaining task is, therefore, to prove the piecewise continuity of  $g(t)$  on  $[-x, \lambda - x]$ , which is actually verified through the following discussion.

When  $t \neq 0$ ,  $f(t)$  is piecewise continuous and  $\sin(t/2)$  and  $e^{it/2}$  are bounded; thus  $g(t)$  is surely piecewise continuous. When  $t = 0$ , we have

$$g(t) = \left\{ \frac{f(x+t) - f(x+0)}{t} + \frac{f(x-t) - f(x-0)}{t} \right\} \cdot \frac{t}{2 \sin(\frac{1}{2}kt)} \cdot 2e^{i\frac{1}{2}kt},$$

so

$$\lim_{t \rightarrow 0} g(t) = \left\{ \lim_{t \rightarrow 0} \frac{f(t+x) - f(x+0)}{t} + \lim_{t \rightarrow 0} \frac{f(x-t) - f(x-0)}{t} \right\} \cdot 2. \quad (11.54)$$

The first and second terms in (11.54) are the derivatives of  $f(x)$  on the right and left, respectively. Since  $f(t)$  is assumed to be piecewise smooth,  $f'(t)$  is piecewise continuous; thus both terms in (11.54) exist. This indicates that the limit  $\lim_{t \rightarrow 0} g(t)$  exists, so  $g(t)$  is piecewise continuous within the interval  $[-x, \lambda - x]$ .

Consequently, we can conclude from (11.53) that

$$\lim_{N \rightarrow \infty} S_N(x) = \frac{f(x+0) + f(x-0)}{2},$$

which implies the pointwise convergence of the Fourier series to  $[f(x+0) + f(x-0)]/2$ .

### 11.3.5 Gibbs Phenomenon

If a function  $f(x)$  has discontinuities in the defining region, its Fourier series does not reproduce the behavior of  $f(x)$  at points of discontinuity. In other words, the partial sums of a Fourier series cannot approach  $f(x)$  uniformly in the vicinity of a point of discontinuity. Furthermore, close to discontinuous points, the Fourier series inevitably overshoots the value of the original function to be expanded. The size of the overshoot is proportional to the magnitude of the discontinuity. This overshoot is known, which as the **Gibbs phenomenon** is nicely illustrated with the Fourier series for the step function

$$f(x) = \begin{cases} +1 & \text{for } 0 \leq x \leq \frac{\lambda}{2}, \\ -1 & \text{for } \frac{\lambda}{2} < x < \lambda, \end{cases}$$

which is a periodic square wave with period  $\lambda$ . The complex Fourier coefficient  $c_n$  reads

$$\begin{aligned} c_n &= \frac{1}{\lambda} \left( \int_0^{\lambda/2} e^{-inkx} dx - \int_{\lambda/2}^{\lambda} e^{-inkx} dx \right) \\ &= \frac{1}{2in\pi} (1 - e^{-in\pi})^2 = \begin{cases} 0 & (n = \text{even}) \\ \frac{2}{in\pi} & (n = \text{odd}). \end{cases} \end{aligned}$$

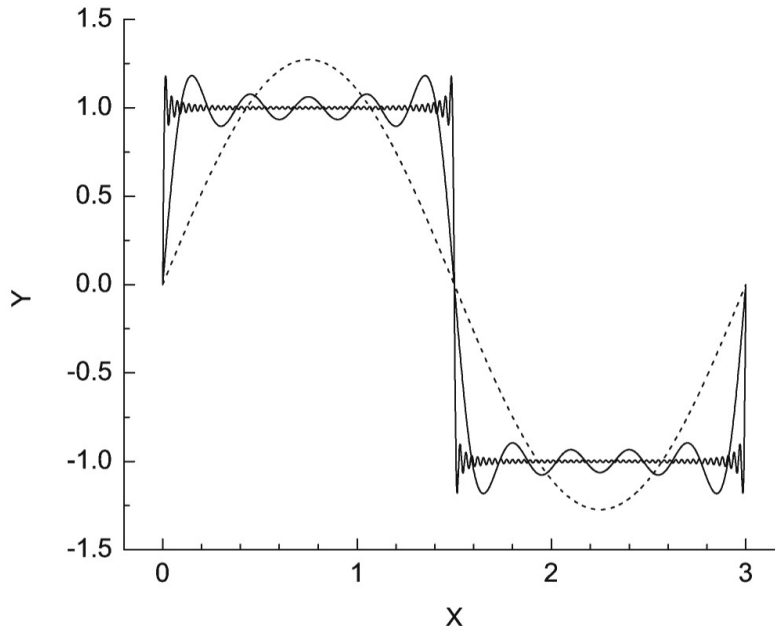
Then we have

$$\begin{aligned} f(x) &\sim \sum_{n=\dots, -3, -1, 1, 3, \dots} \frac{2}{in\pi} e^{inkx} = \sum_{n=1, 3, \dots} \left( \frac{2}{-in\pi} e^{-inkx} + \frac{2}{in\pi} e^{inkx} \right) \\ &= \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{\sin(2n-1)kx}{2n-1}. \end{aligned} \quad (11.55)$$

Figure 11.7 shows  $f(x)$  for  $0 \leq x \leq \lambda$  for the sum of four, six, and ten terms of the series. Three features deserve attention.

- (i) There is a steady increase in the accuracy of the representation as the number of included terms is increased.
- (ii) All the curves pass through the midpoint of  $f(x) = 0$  at the points of discontinuity  $x = n\lambda/2$  ( $n = 0, \pm 1, \pm 2, \dots$ ).
- (iii) In the vicinity of  $x = n\lambda/2$ , there is an overshoot that persists and shows no sign of diminishing.

As more and more terms are taken, the small oscillations along each horizontal portion get smaller and smaller and, except for the two outer terms of



**Fig. 11.7.** Gibbs phenomena for the Fourier series of a step function. The partial sums of one, five, and fifty terms the right-hand side of (11.55) are given

each portion closes to the discontinuities, eventually disappear. Even in the limit of an infinite number of terms, there is still a small overshoot. This overshoot is nothing but what we call the **Gibbs phenomenon**, which results in the fact that the Fourier series cannot have uniform convergence at a point of discontinuity.

### 11.3.6 Overshoot at a Discontinuous Point

Owing to Gibbs phenomena, a Fourier series representation is highly unreliable in the vicinity of a discontinuity. We now consider the resulting degree of error when we represent a function  $f(x)$  by a Fourier series having a discontinuity.

The maximum overshoot can be evaluated analytically through the following procedure. Let us consider a finite sum of the Fourier series in the complex form

$$S_N(x) = \sum_{n=-N}^N c_n e^{inkx},$$

which yields

$$S_N(x) = \frac{1}{\lambda} \int_{-x}^{\lambda-x} f(t+x) K_N(t) dt, \quad K_N(t) \equiv \frac{\sin[(N + \frac{1}{2})kt]}{\sin(\frac{1}{2}kt)}. \quad (11.56)$$

We consider the behavior of  $S_N(x)$  in the vicinity of a discontinuity at  $x = x_0$ . We denote the jump of  $f(x)$  at this discontinuity by  $\Delta f$  and the jump of its finite Fourier sum by  $\Delta S_N$ :

$$\Delta f \equiv f(x_0 + \varepsilon) - f(x_0 - \varepsilon), \quad \Delta S_N \equiv S_N(x_0 + \varepsilon) - S_N(x_0 - \varepsilon),$$

where  $\varepsilon$  is infinitesimal. We then have

$$\Delta S_N = \frac{1}{\lambda} \int_{-x_0-\varepsilon}^{\lambda-x_0-\varepsilon} f(t+x_0+\varepsilon) K_N(t) dt - \frac{1}{\lambda} \int_{-x_0+\varepsilon}^{\lambda-x_0+\varepsilon} f(t+x_0-\varepsilon) K_N(t) dt.$$

Owing to the periodicity of the integrand  $f(t+x)K_N(t)$ , we replace the range of integration as follows:

$$\Delta S_N = \frac{1}{\lambda} \int_{-\varepsilon}^{\lambda-\varepsilon} f(t+x_0+\varepsilon) K_N(t) dt - \frac{1}{\lambda} \int_{\varepsilon}^{\lambda+\varepsilon} f(t+x_0-\varepsilon) K_N(t) dt.$$

Hence, we have

$$\begin{aligned} \Delta S_N &= \frac{1}{\lambda} \left( \int_{-\varepsilon}^{\varepsilon} + \int_{\varepsilon}^{\lambda-\varepsilon} \right) f(t+x_0+\varepsilon) K_N(t) dt \\ &\quad - \frac{1}{\lambda} \left( \int_{\varepsilon}^{\lambda-\varepsilon} + \int_{\lambda-\varepsilon}^{\lambda+\varepsilon} \right) f(t+x_0-\varepsilon) K_N(t) dt \\ &= \frac{1}{\lambda} \int_{-\varepsilon}^{\varepsilon} [f(t+x_0+\varepsilon) - f(t+x_0-\varepsilon)] K_N(t) dt \\ &\quad + \frac{1}{\lambda} \int_{\varepsilon}^{\lambda-\varepsilon} [f(t+x_0+\varepsilon) - f(t+x_0-\varepsilon)] K_N(t) dt. \end{aligned} \quad (11.57)$$

The integrand of (11.57) gives zero for all values of  $t$  except near  $t = 0$ . Close to  $t = 0$ , the integrand has a somewhat large value because of (i) the jump of  $f(t+x_0)$  at  $t = 0$  and (ii) the significant contribution of  $K_N(t)$  in the vicinity of  $t = 0$ . Hence, we can confine the integration to the small interval  $(-\delta, +\delta)$  for which the difference in the square brackets in (11.57) is simply  $\Delta f$ . It now follows that

$$\Delta S_N \simeq \frac{\Delta f}{\lambda} \int_{-\delta}^{\delta} \frac{\sin \left\{ \left( N + \frac{1}{2} \right) kt \right\}}{\sin \frac{1}{2} kt} dt \simeq \frac{4\Delta f}{\lambda} \int_0^{\delta} \frac{\sin \left\{ \left( N + \frac{1}{2} \right) kt \right\}}{kt} dt, \quad (11.58)$$

where the sine in the dominator was approximated by its argument because of the smallness of  $t$ .

The value of  $\Delta S_N$  depends crucially on the interval  $\delta$ , since the integrand in (11.58) rapidly alternates its sign as  $t$  increases. The reader may find the plot of the integrand in Fig. 11.8, where it is shown clearly that the major contribution to the integral comes from the interval  $[0, \lambda/(2N+1)]$ , where  $\lambda/(2N+1)$  is the first zero of the integrand. Hence, if the upper limit is larger than  $\lambda/(2N+1)$ , the result of the integral will clearly decrease, because in each interval of length  $\lambda$ , the area below the horizontal axis is larger than that above. Therefore, if we are interested in the maximum overshoot of the finite

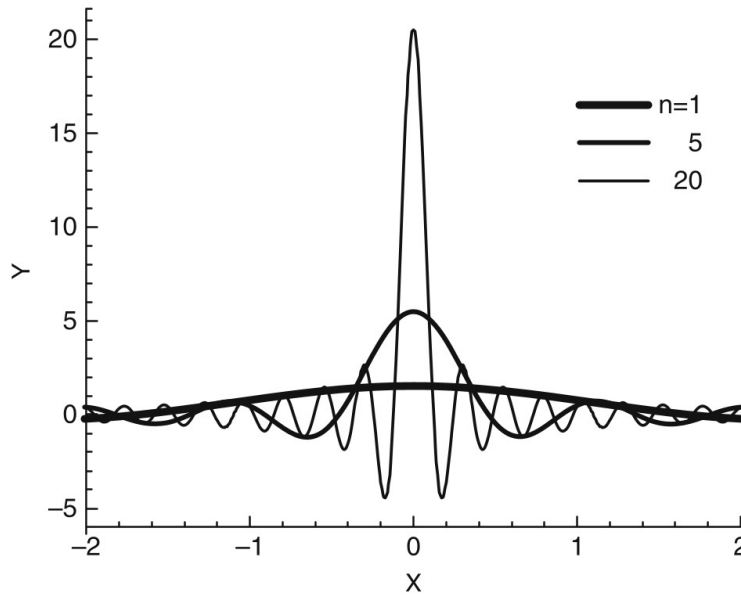


Fig. 11.8. The integrand of (11.58)

sum  $\Delta S_N$ , we must set the upper limit equal to  $\lambda/(2N+1)$ . It follows that the maximum overshoot is

$$\begin{aligned}
 (\Delta S_N)_{\max} &\simeq \frac{4\Delta f}{\lambda} \int_0^{\lambda/(2N+1)} \frac{\sin(N + \frac{1}{2})kt}{kt} dt \\
 &= \frac{4\Delta f}{\lambda} (N + \frac{1}{2}) \int_0^{\pi} \frac{\sin x}{x} \frac{dx}{(N + \frac{1}{2})k} = \frac{2\Delta f}{\pi} \int_0^{\pi} \frac{\sin x}{x} dx \\
 &\simeq 1.179\Delta f.
 \end{aligned}$$

We thus conclude that the finite (large- $N$ ) sum approximation of the discontinuous function overshoots the function itself at a discontinuity by about 18% in this case. This means that the Fourier series tends to overshoot the positive corner by some 18% and to undershoot the negative corner by the same amount. The inclusion of more terms (increasing  $r$ ) does nothing to remove this overshoot but merely moves it closer to the point of discontinuity.

## Exercises

1. Let  $f(x)$  be absolutely integrable and form the Fourier series of  $f(x)$  in the interval  $(-\pi, \pi)$ . Show that the convergence of its Fourier series at a specified point  $x$  within the interval depends only on the behavior of  $f$  in the immediate vicinity of this point. (This result is referred to as the **localization theorem**.)

**Solution:** We use the integral formula for the partial sums

$$\begin{aligned}
S_n(x) &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(x+u) \frac{\sin mu}{2 \sin(u/2)} du \\
&= \frac{1}{\pi} \int_{-\delta}^{\delta} f(x+u) \frac{\sin mu}{2 \sin(u/2)} du + I_1 + I_2,
\end{aligned}$$

where we have set  $m = n + (1/2)$ . Here  $\delta$  is an arbitrarily small positive number, and  $I_1, I_2$  are the integrals over the intervals  $[\delta, \pi]$  and  $[-\pi, -\delta]$ , respectively. On these intervals, the function  $1/[2 \sin(u/2)]$  is continuous (since  $|u| \geq \delta$ ) and, therefore, the function

$$\phi(u) = \frac{f(x+u)}{2 \sin(u/2)}$$

is absolutely integrable. It then follows from the Riemann–Lebesgue theorem that the integral

$$I_1 = \frac{1}{\pi} \int_{\delta}^{\pi} \phi(u) \sin mudu$$

approaches zero as  $m \rightarrow \infty$ . The same is true of  $I_2$ . Thus, whether or not the partial sums of the Fourier series have a limit at the point  $x$  depends on the behavior of the integral

$$\frac{1}{\pi} \int_{-\delta}^{\delta} f(x+u) \frac{\sin mu}{2 \sin(u/2)} du$$

as  $m \rightarrow \infty$ , which involves only the values of the function  $f(x)$  in the neighborhood  $[x - \delta, x + \delta]$  of the point  $x$ . This completes the proof. ♣

2. Let  $f(x) = -\log |2 \sin(x/2)|$ , which is even and becomes infinite at  $x = 2k\pi$  ( $k = 0, \pm 1, \pm 2, \dots$ ).

- (i) Show that  $f(x)$  is integrable.
- (ii) Calculate the Fourier series of  $f(x)$ .
- (iii) Derive the identity:  $\log 2 = 1 - (1/2) + (1/3) - (1/4) + \dots$ .

**Solution:**

- (i) The given  $f(x)$  equals zero at  $x = \pi/3$  and is  $2\pi$ -periodic. Hence, to prove the integrability of  $f(x)$ , it suffices to show that it is integrable on the interval  $[0, \pi/3]$ . Clearly we have

$$-\int_{\varepsilon}^{\pi/3} \log \left| 2 \sin \frac{x}{2} \right| dx = \varepsilon \log \left( 2 \sin \frac{\varepsilon}{2} \right) + \int_{\varepsilon}^{\pi/3} \frac{x \cos(x/2)}{2 \sin(x/2)},$$

where we have dropped the absolute value sign, since  $2 \sin(x/2) > 1$  for  $0 < x < \pi/3$ . As  $\varepsilon \rightarrow 0$ , the quantity  $\varepsilon \log[2 \sin(\varepsilon/2)]$  approaches zero, which is verified by using **l'Hôpital's rule** (see Sect. 1.4.1), whereas the last integral converges since the integrand is bounded. (Recall that  $\lim_{x \rightarrow 0} x/[2 \sin(x/2)] = 1$ .) Thus,  $-\int_0^{\pi/3} \log |2 \sin \frac{x}{2}| dx$  exists, i.e.,  $f(x)$  is integrable on the interval  $[0, \pi/3]$ .

(ii) Since  $f(x)$  is even, we have  $b_n = 0$  ( $n = 1, 2, \dots$ ) and

$$a_n = -\frac{2}{\pi} \int_0^\pi \log \left( 2 \sin \frac{x}{2} \right) \cos nx dx \quad (n = 0, 1, 2, \dots).$$

For  $n \neq 0$ , integrating by parts and then applying l'Hôpital's rule, we get

$$a_n = \frac{1}{n\pi} \int_0^\pi \frac{\sin nx \cos(x/2)}{\sin(x/2)} dx \quad (n = 1, 2, \dots),$$

and then use the identity  $2 \sin nx \cos(x/2) = \sin[n + (1/2)]x + \sin[n - (1/2)]x$  to obtain

$$\begin{aligned} a_n &= \frac{1}{n\pi} \int_0^\pi \frac{\sin[n + (1/2)]x}{2 \sin(x/2)} dx + \frac{1}{n\pi} \int_0^\pi \frac{\sin[n - (1/2)]x}{2 \sin(x/2)} dx \\ &= \frac{1}{n}. \quad (n = 0, 1, 2, \dots). \end{aligned}$$

For  $n = 0$ , we have

$$\begin{aligned} a_0 &= -\frac{2}{\pi} \int_0^\pi \log \left( 2 \sin \frac{x}{2} \right) dx = -\frac{2}{\pi} \int_0^\pi \left( \log 2 + \log \sin \frac{x}{2} \right) dx \\ &= \pi \log 2 + \int_0^\pi \log \left( \sin \frac{x}{2} \right) dx. \end{aligned}$$

The last integral, denoted by  $I$ , reads

$$\begin{aligned} I &= 2 \int_0^{\pi/2} \log(\sin t) dt = 2 \int_0^{\pi/2} \log \left( 2 \sin \frac{t}{2} \cos \frac{t}{2} \right) dt \\ &= \pi \log 2 + 2 \int_0^{\pi/2} \log \left( \sin \frac{t}{2} \right) dt + 2 \int_0^{\pi/2} \log \left( \cos \frac{t}{2} \right) dt. \end{aligned}$$

The substitution  $t = \pi - u$  gives  $\int_0^{\pi/2} \log[\cos(t/2)] dt = \int_{\pi/2}^\pi \log[\sin(u/2)] du$ , which implies that  $I = \pi \log 2 + 2I$ , i.e.,  $I = -\pi \log 2$ . Consequently,  $a_0 = 0$ .

- (iii) Since the function  $f(x)$  is obviously differentiable for  $x \neq 2k\pi$  ( $k = 0, \pm 1, \pm 2, \dots$ ), it follows that

$$-\log \left| 2 \sin \frac{x}{2} \right| = \cos x + \frac{\cos 2x}{2} + \frac{\cos 3x}{3} + \dots \quad (11.59)$$

for  $x \neq 2k\pi$  ( $k = 0, \pm 1, \pm 2, \dots$ ). Setting  $x = \pi$  in (11.59), we obtain the desired result. ♣

3. Show that

$$\int_{-x}^{\lambda-x} \frac{e^{i(N+\frac{1}{2})kt}}{\sin(\frac{1}{2}kt)} dt = i\lambda. \quad (11.60)$$

**Solution:**

Recall an alternative form of  $S_N(x)$  given in (11.40):

$$S_N(x) = \frac{1}{\lambda} \int_{-x}^{\lambda-x} f(t+x) \left( \sum_{n=-N}^N e^{-inkt} \right) dt. \quad (11.61)$$

Setting  $f(t) \equiv 1$  into (11.61) and (11.52) and comparing them, we have

$$\begin{aligned} 0 + \frac{1}{i\lambda} \int_{-x}^{\lambda-x} \frac{e^{i\frac{1}{2}kt} e^{iNkt}}{\sin(\frac{1}{2}kt)} dt &= \frac{1}{\lambda} \int_{-x}^{\lambda-x} \left( \sum_{n=-N}^N e^{-inkt} \right) dt \\ &= \frac{1}{\lambda} \sum_{n=-N}^N \left( \int_{-x}^{\lambda-x} e^{-inkt} dt \right) = \frac{1}{\lambda} \sum_{n=-N}^N \lambda \delta_{n,0} = 1. \quad \clubsuit \end{aligned}$$

## 11.4 Applications in Physics and Engineering

### 11.4.1 Temperature Variation of the Ground

The most important applications of Fourier series expansions in the physical sciences are in solving **partial differential equations** that describe a wide variety of physical phenomena. In this section, two typical examples of such applications are presented, while more rigorous discussions on partial differential equations are given in Chap. 17.

First, we consider the temperature variation of the ground exposed to sunlight. The temperature at a depth of  $x$  meters at time  $t$ , denoted by  $u(x, t)$ , is known to be determined by the **diffusion equation**

$$\frac{\partial u}{\partial t} = \kappa \frac{\partial^2 u}{\partial x^2}. \quad (11.62)$$

Here, the proportionality constant  $\kappa$  is called the **thermal conductivity** and its magnitude on the ground is roughly estimated at  $\kappa = 3.0 \times 10^{-6} \text{ m}^2/\text{s}$ . We will see below that the Fourier series expansion provides a means of solving equation (11.62) and clarifying the physical interpretation of its solution.

Suppose that the temperature of the land surface,  $u(x = 0, t)$ , changes periodically with a period  $T$ ; the period  $T$  may range from a day to a year. It is then reasonable to express  $u(x, t)$  by the Fourier series

$$u(x, t) = \sum_{n=-\infty}^{\infty} c_n(x) e^{in\omega t} \quad \left( \omega = \frac{2\pi}{T} \right).$$

Substitute this into (11.62) to obtain

$$in\omega c_n(x) = \kappa \frac{d^2 c_n}{dx^2},$$

which implies

$$c_n(x) \propto \begin{cases} \exp\left(-\frac{1+i}{\sqrt{2}} \sqrt{\frac{n\omega}{\kappa}} x\right) & n > 0 \\ \exp\left(-\frac{1-i}{\sqrt{2}} \sqrt{\frac{|n|\omega}{\kappa}} x\right) & n < 0. \end{cases}$$

Here we have chosen the solutions that behave as  $|c_n(x)| \rightarrow 0$  in the limit of  $x \rightarrow \infty$ . In order to obtain the zeroth term  $c_0(x)$ , we note that

$$\frac{dc_0(x)}{dx^2} = 0,$$

and thus

$$c_0(x) = A_0 + B_0 x.$$

Owing to the condition that  $\lim_{x \rightarrow \infty} |c_0(x)| = 0$ , we see that  $B_0 = 0$  and  $A_0 = \text{const}$ . As a result, we obtain

$$u(x, t) = A_0 + 2 \sum_{n=1}^{\infty} A_n e^{-\alpha_n x} \cos(n\omega t - \alpha_n x + \phi_n), \quad (11.63)$$

where

$$\alpha_n = \sqrt{\frac{n\omega}{2\kappa}}$$

and the constants  $A_n$  and  $\phi_n$  are determined by the  $t$ -dependence of the surface temperature  $u(x = 0, t)$ .

Note the presence of the parameter  $\alpha_n$  in the general solution (11.63). It indicates that a wave component with the period  $T/n$  has the following

features: (i) decay of the wave amplitude by  $e^{-\alpha_n x}$  with an increase in  $x$ , and (ii) a phase shift by  $\alpha_n x$  relative to the surface temperature  $u(x = 0, t)$ .

Let us quantify the actual value of  $\alpha_n$ . For this, we consider the case of  $T = 1$  day (i.e.,  $60 \times 60 \times 24$  s) and assume monochromatic variation of the surface temperature given by

$$u(0, t) = 15 + 5 \cos\left(\frac{2\pi}{T}t\right) \text{ } ^\circ\text{C}.$$

Comparing this with (11.63) with  $x = 0$ , we get  $A_0 = 15$ ,  $A_1 = 5/2$ , and  $A_n = 0$  for  $n \geq 2$ . Then, since

$$\alpha_1 = \sqrt{\frac{2 \times 3.14}{2 \times (3.0 \times 10^{-6}) \times (60 \times 60 \times 24)}} \simeq 3.5,$$

we have

$$u(x, t) = 15 + 5e^{-3.5x} \cos\left(\frac{2\pi}{T}t - 3.5x\right).$$

A three-dimensional plot of  $u(x, t)$  in the  $x$ - $t$  plane is shown in Fig. 11.9. We observe that at depths greater than 1 m, the temperature variation is almost in antiphase to that at the surface ( $x = 0$ ) and the amplitude decreases considerably.

#### 11.4.2 String Vibration Under Impact

The second example is the vibration of an elastic string subject to an impact force in a local region. Consider the case of a piano wire under an impact force applied by a hammer. Suppose that an impulse  $I$  is applied at the position  $x = a$  of a suspended string with length  $\ell$  and mass density  $\rho$ . The vibrational amplitude of the string, denoted by  $u(x, t)$ , is governed by the **wave equation**

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}. \quad (11.64)$$

The string is initially assumed to be stationary, i.e.,

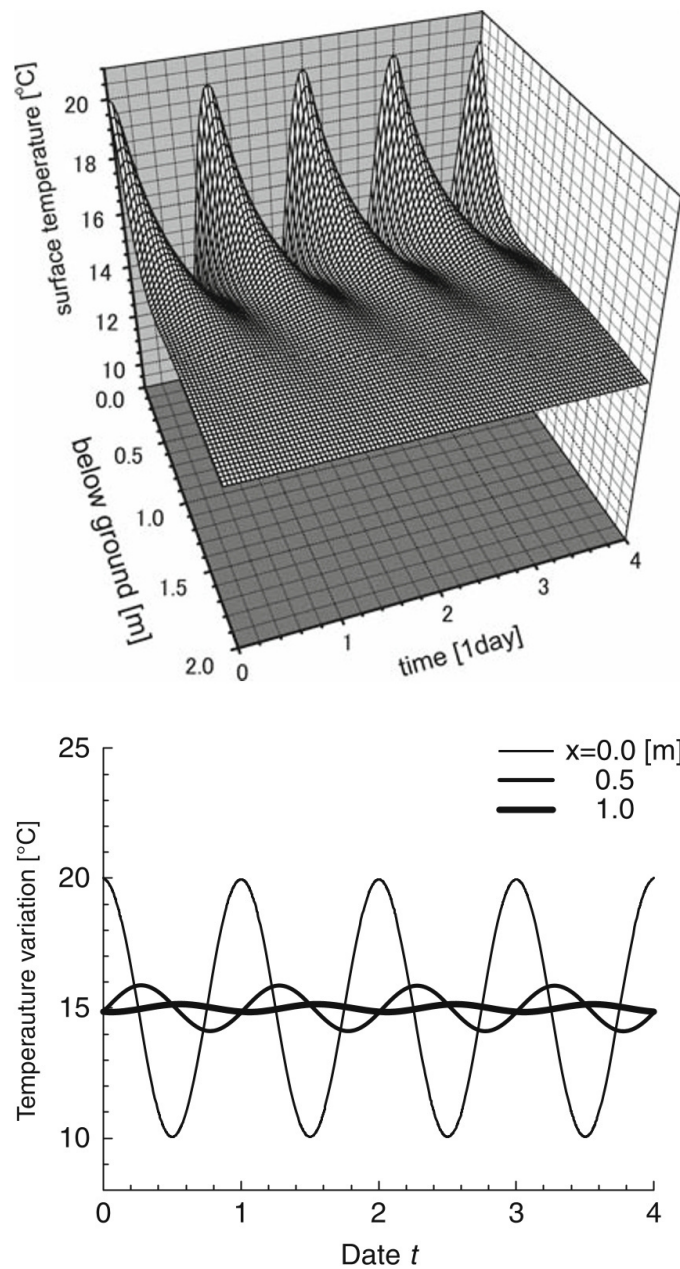
$$u(x, t = 0) = 0. \quad (11.65)$$

The initial velocity of the line element at  $x$  is denoted by  $v(x)$ . Then, the **law of the conservation of momentum** states that

$$\int_0^\ell \rho v(x) dx = I, \quad (11.66)$$

where

$$v(x) = V \cdot \delta(x - a), \quad (11.67)$$



**Fig. 11.9.** Temperature variation  $u(x, t)$  of the underground below  $x$  meters on  $t$  days

with an appropriate constant  $V$ . From (11.66) and (11.67), we have  $V = I/\rho$ . Furthermore, since

$$v(x) = \left. \frac{\partial u}{\partial t} \right|_{t=0},$$

we have

$$\left. \frac{\partial u}{\partial t} \right|_{t=0} = \frac{I}{\rho} \delta(x - a). \quad (11.68)$$

Under the two initial conditions (11.65) and (11.68), the general solution of (11.64) is given by

$$u(x, t) = \sum_{n=1}^{\infty} A_n \sin k_n x \sin (\omega_n t + \phi_n), \quad (11.69)$$

where

$$k_n = \frac{n\pi}{\ell}, \quad \omega_n = ck_n.$$

The constants  $A_n$  and  $\phi_n$  in (11.69) are again determined by the initial conditions. First, imposing the condition  $u(x, t = 0) = 0$  into (11.69) implies

$$A_n \sin \phi_n = 0 \quad \text{for all } n, \quad (11.70)$$

owing to the linear independence of  $\{\sin k_n x\}$ . Next, it follows from (11.69) that

$$\left. \frac{\partial u}{\partial t} \right|_{t=0} = \sum_{n=1}^{\infty} A_n \omega_n \cos \phi_n \sin k_n x = \frac{I}{\rho} \delta(x - a).$$

Multiplying both sides by  $\sin k_m x$  and then integrating yields

$$A_m \omega_m \cos \phi_m \int_0^{\ell} \sin^2 k_m x dx = \frac{I}{\rho} \sin k_m a \quad \text{for all } m. \quad (11.71)$$

From (11.70) and (11.71), we finally obtain

$$\phi_n = 0 \quad \text{and} \quad A_n = \frac{2I}{\rho \ell \omega_n} \sin k_n a \quad \text{for all } n. \quad (11.72)$$

The second expression in (11.72) implies that the position  $x = a$  that satisfies  $\sin k_n a = 0$  yields  $A_n = 0$ ; i.e., the  $n$ th vibration mode is not excited by the impulsive force applied at  $x = a$  that satisfies  $\sin k_n a = 0$ . In contrast, if we apply an impulsive force at  $x = a$  satisfying  $|\sin k_n a| = 1$ , the corresponding  $n$ th mode will have a large vibrational amplitude, as is actually the case inside a piano.



## Fourier Transformation

---

**Abstract** Fourier transformation is an effective tool for confirming the dual nature of a complex-valued function (as well as a real-valued one). Furthermore, the transformation enables us to measure certain correlations of a function with itself or with other functions; thus a Fourier transform can be applied to probability theory, signal analysis, etc. In this chapter we also provide the essence of a discrete Fourier transform (Sect. 12.3), which refers to a Fourier transform applied to a discrete complex-valued series. A discrete Fourier transform is commonly used in the numerical computation of Fourier transforms because of its computational efficiency.

### 12.1 Fourier Transform

#### 12.1.1 Derivation of Fourier Transform

The properties of Fourier series that we have already developed are adequate for handling the expansion of any periodic function. Nevertheless, there are many problems in physics and engineering that do not involve periodic functions, so it is important to generalize Fourier series to include nonperiodic functions. A nonperiodic function can be considered as a limit of a given periodic function whose period becomes infinite.

Let us write Fourier series representing a periodic function  $f(x)$  in complex form:

$$f(x) = \sum_{n=-\infty}^{\infty} c_n e^{ikx}, \quad (12.1)$$

with the definition  $k = 2n\pi/\lambda$ , in which

$$c_n = \frac{1}{\lambda} \int_{-\frac{\lambda}{2}}^{\frac{\lambda}{2}} f(x) e^{-ikx} dx.$$

We then introduce the quantity

$$\Delta k = \frac{2\pi}{\lambda} \Delta n. \quad (12.2)$$

From the definition (12.2), the adjacent values of  $k$  are obtained by setting  $\Delta n = 1$ , which corresponds to  $(\lambda/2\pi)\Delta k = 1$ . Therefore, multiplying each side of (12.1) by  $(\lambda/2\pi)\Delta k$  yields

$$f(x) = \sum_{n=-\infty}^{\infty} c_{\lambda}(k) e^{ikx} \Delta k, \quad (12.3)$$

where

$$c_{\lambda}(k) \equiv \frac{\lambda}{2\pi} c_n = \frac{1}{2\pi} \int_{-\frac{\lambda}{2}}^{\frac{\lambda}{2}} f(x) e^{-ikx} dx.$$

In the limit as  $\lambda \rightarrow \infty$ , the  $k$ s are distributed continuously instead of discretely, i.e.,  $\Delta k \rightarrow dk$ . Thus, the sum in (12.3) becomes exactly the definition of an integral. As a result, we arrive at the conclusion

$$c(k) = \lim_{\lambda \rightarrow \infty} c_{\lambda}(k) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(x) e^{-ikx} dx \quad (12.4)$$

and

$$f(x) = \int_{-\infty}^{\infty} c(k) e^{ikx} dk. \quad (12.5)$$

Further, by defining  $F(k) = \sqrt{2\pi}c(k)$ , equations (12.4) and (12.5) take the symmetrical form given below, known as the **Fourier transform** or **Fourier integral representation** of  $f(x)$ .

♠ **Fourier transform:**

The Fourier transform of  $f(x)$  is defined by

$$F(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-ikx} dx. \quad (12.6)$$

♠ **Inverse Fourier transform:**

The inverse Fourier transform of  $F(k)$  given above is defined by

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} F(k) e^{ikx} dk, \quad (12.7)$$

We often write the expressions (12.6) and (12.7) in simpler form:

$$F(k) = \mathcal{F}[f(x)] \quad \text{and} \quad f(x) = \mathcal{F}^{-1}[F(k)].$$

Observe that  $F(k)$  as well as  $f(x)$  are, in general, complex-valued functions of the real variables  $k$  and  $x$ , respectively. Yet, if  $f(x)$  is real, then

$$F(-k) = F^*(k),$$

which gives two immediate corollaries (proofs are left to the reader):

♠ **Fourier integral theorem:**

1. If  $f(x)$  is real and even,  $F(k)$  is real.
2. If  $f(x)$  is real and odd,  $F(k)$  is purely imaginary.

### 12.1.2 Fourier Integral Theorem

Our derivations of the Fourier transform and its inverse transform, (12.7) and (12.6), have been ambiguous from a mathematical viewpoint. For developing exact derivations and clarifying the conditions for the infinite integrals in (12.7) and (12.6) to converge, the following theorem is of crucial importance:

♠ **Fourier integral theorem:**

If  $f(x)$  is piecewise smooth and absolutely integrable, then

$$\frac{1}{\pi} \int_0^\infty \left[ \int_{-\infty}^\infty f(t) \cos u(x-t) dt \right] du = \frac{f(x+0) + f(x-0)}{2}. \quad (12.8)$$

**Remark.** The theorem is valid for each fixed  $x$ , so  $x$  can be considered a constant insofar as the integrations are concerned.

Before starting the proof of the theorem, we note that (12.8) reduces to the form of (12.7) and (12.6) when  $x$  is a continuous point of  $f(x)$ . To see this, we make use of the identity

$$\int_0^\xi \cos u(x-t) du = \frac{1}{2} \int_{-\xi}^\xi e^{iu(x-t)} du. \quad (12.9)$$

Since (12.8) reads

$$f(x) = \lim_{\xi \rightarrow \infty} \frac{1}{\pi} \int_{-\infty}^\infty f(t) dt \int_0^\xi \cos u(x-t) du, \quad (12.10)$$

we substitute (12.9) and (12.10) to obtain

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^\infty e^{iux} du \int_{-\infty}^\infty f(t) e^{-itu} dt = \frac{1}{2\pi} \int_{-\infty}^\infty F(u) e^{iux} du,$$

where

$$F(u) = \frac{1}{2\pi} \int_{-\infty}^\infty f(t) e^{-itu} dt.$$

These results are clearly equivalent to the forms of (12.7) and (12.6).

### 12.1.3 Proof of the Fourier Integral Theorem

The proof of the Fourier integral theorem is based on the following two lemmas:

♠ **Lemma 1:** If  $f(x)$  is piecewise smooth for all  $x \in \mathbf{R}$ , then

$$\lim_{\xi \rightarrow \infty} \int_0^b f(x) \frac{\sin \xi x}{x} dx = \frac{\pi}{2} f(0+) \quad \text{for } b > 0.$$

♠ **Lemma 2:** If  $f(x, t)$  is a continuous function of  $t$  for  $a \leq t \leq b$  and if  $\lim_{c \rightarrow \infty} \int_0^c f(x, t) dx$  exists and converges uniformly to a certain function  $g(t)$  in the interval, then  $g(t)$  is continuous in the interval and

$$\int_a^b g(t) dt = \int_a^b \left[ \int_0^\infty f(x, t) dx \right] dt = \int_0^\infty \left[ \int_a^b f(x, t) dt \right] dx.$$

Note that  $f(0+)$  in Lemma 1 denotes the limiting value of  $f(x)$  as  $x$  tends to zero through positive values. The proof of Lemma 1 is left to Exercise 2. Lemma 2 follows from the fact that uniform convergence allows us to interchange the order of limiting and integration procedures (see Chapter 3 for details).

We are now ready to prove the Fourier integral theorem expressed by (12.8).

**Proof (of the Fourier integral theorem):** Let  $f(x)$  be piecewise smooth and absolutely integrable. Consider the integral

$$\int_{-\infty}^{\infty} f(t) \cos u(x - t) dt.$$

Since  $|\cos u(x - t)| \leq 1$ , the convergence of this integral is ensured by our hypothesis that  $\int_{-\infty}^{\infty} |f(t)| dt$  converges, and since this conclusion is independent of  $u$  and  $x$ , the convergence is uniform for all  $u$ . Therefore, in view of Lemma 2, we can interchange the order of integration in

$$I \equiv \int_0^b \left[ \int_{-\infty}^{\infty} f(t) \cos u(x - t) dt \right] du$$

to obtain

$$I \equiv \int_{-\infty}^{\infty} \left[ \int_0^b f(t) \cos u(x - t) du \right] dt = \int_{-\infty}^{\infty} \frac{\sin b(x - t)}{x - t} f(t) dt.$$

We now decompose this into four integrals:

$$I = \left[ \int_{-\infty}^{-M} + \int_{-M}^x + \int_x^M + \int_M^{\infty} \right] \frac{\sin b(x - t)}{x - t} f(t) dt, \quad (12.11)$$

where  $M$  is taken to be so large that the first and the last integrals in (12.11) are less in absolute value than some prescribed  $\varepsilon > 0$ . By changing variables, taking  $u = t - x$ , we can write the third integral in (12.11) as

$$\int_0^{M-x} \frac{\sin bu}{u} f(x+u) du.$$

In view of Lemma 1, this tends to  $\pi f(x+0)/2$  as  $b \rightarrow \infty$ . Similarly, the second integral tends to  $\pi f(x-0)/2$ . Therefore, by taking  $M$  sufficiently large, we obtain

$$\lim_{b \rightarrow \infty} I < \frac{\pi [f(x+0) + f(x-0)]}{2} + 2\varepsilon,$$

or equivalently,

$$\int_0^\infty \left[ \int_{-\infty}^\infty f(t) \cos u(x-t) dt \right] du - \frac{\pi [f(x+0) + f(x-0)]}{2} < 2\varepsilon.$$

This completes the proof of the theorem. ♣

#### 12.1.4 Inverse Relations of the Half-width

In practice, we often encounter functions  $f(x)$  having a sharp peak at a specific point, say  $x = 0$ . The width of the peak of such a function is possibly correlated with the width of the peak that is exhibited by the resulting Fourier transform  $F(k) = \mathcal{F}[f(x)]$ . A typical example of this phenomenon is seen by considering the Fourier transform of a Gaussian function  $f(x) = ae^{-bx^2}$  with  $a, b > 0$ , i.e.,

$$F(k) = \frac{a}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-bx^2} e^{-ikx} dx = \frac{ae^{-k^2/(4b)}}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-b[x+ik/(2b)]^2} dx.$$

We substitute  $y = x + ik/(2b)$  to evaluate the integral as

$$\int_{-\infty}^\infty e^{-b[x+ik/(2b)]^2} dx = \int_{-\infty}^\infty e^{-by^2} dy = \sqrt{\frac{\pi}{b}}.$$

and we get

$$F(k) = \frac{a}{\sqrt{2b}} e^{-k^2/(4b)},$$

which is also Gaussian. It is noteworthy that the width of  $f(x)$ , which is proportional to  $1/\sqrt{b}$ , is in inverse relation to the width of  $F(k)$ , which is proportional to  $\sqrt{b}$ . Therefore, increasing the width of  $f(x)$  results in a decrease in the width of  $F(k)$ . In the limit of infinite width (a constant function), we get infinite sharpness (the delta function). In fact, denoting the widths as  $\Delta x$  and  $\Delta k$ , we have  $\Delta x \Delta k \sim 1$ .

♠ **Inverse relation of the half-width:**

When  $f(x)$  consists of a single peak whose width is characterized by  $\Delta x$ , its Fourier transform  $F(k)$  is also a single-peak function with a width  $\Delta k$ , which yields  $\Delta x \Delta k \sim 1$ .

For the second example, we evaluate the Fourier transform of a box function defined by

$$f(x) = \begin{cases} b, & \text{if } |x| < a, \\ 0, & \text{if } |x| > a. \end{cases}$$

From the definition, we have

$$F(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-ikx} dx = \frac{b}{\sqrt{2\pi}} \int_{-a}^a e^{-ikx} dx = \frac{2ab}{\sqrt{2\pi}} \left( \frac{\sin ka}{ka} \right).$$

Observe again that the width of  $f(x)$ ,  $\Delta x = 2a$ , is in inverse relation to the width of  $F(k)$ , which is roughly the distance between its first two roots,  $k_+$  and  $k_-$ , on either side of  $k = 0$ :  $\Delta k = k_+ - k_- = 2\pi/a$ . In addition, if  $a \rightarrow \infty$ , the function  $f(x)$  becomes a constant function over the entire real line, and we get

$$F(k) = \frac{2b}{\sqrt{2\pi}} \lim_{a \rightarrow \infty} \frac{\sin ka}{k} = \frac{2b}{\sqrt{2\pi}} \pi \delta(k).$$

Otherwise, if  $b \rightarrow \infty$  and  $a \rightarrow 0$  in such a way that  $2ab$  [the area under the graph of  $f(x)$ ] remains fixed at unity, then  $f(x)$  approaches the delta function and  $F(k)$  becomes

$$F(k) = \lim_{a \rightarrow 0} \lim_{b \rightarrow \infty} \frac{2ab}{\sqrt{2\pi}} \frac{\sin ka}{ka} = \frac{1}{\sqrt{2\pi}}.$$

### 12.1.5 Parseval Identity for Fourier Transforms

If  $F(k)$  and  $G(k)$  are Fourier transforms of  $f(x)$  and  $g(x)$ , respectively, we have

$$\begin{aligned} & \int_{-\infty}^{\infty} f(x) g^*(x) dx \\ &= \int_{-\infty}^{\infty} \left\{ \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} F(k) e^{-ikx} dk \right\} \times \left\{ \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} G^*(k') e^{ik'x} dk' \right\} dx \\ &= \int_{-\infty}^{\infty} dk \int_{-\infty}^{\infty} dk' F(k) G^*(k') \left\{ \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-i(k-k')x} dx \right\} \\ &= \int_{-\infty}^{\infty} dk F(k) \int_{-\infty}^{\infty} dk' G^*(k') \delta(k' - k) \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} F(k) G^*(k) dk, \end{aligned} \tag{12.12}$$

or similarly,

$$\int_{-\infty}^{\infty} f(x)g(x) dx = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(k)G(-k) dk. \quad (12.13)$$

In particular, if we set  $g(x) = f(x)$  in (12.12), we have

$$\int_{-\infty}^{\infty} |f(x)|^2 dx = \frac{1}{2\pi} \int_{-\infty}^{\infty} |F(k)|^2 dk. \quad (12.14)$$

Here  $|F(k)|^2$  is referred to as the **power spectrum** of the function  $f(x)$ . Equation (12.14), or the more general (12.13), is known as the **Parseval identity** for Fourier integrals.

**Remark.** A sufficient condition for interchanging the order of integration in (12.12) is the absolute convergence of the integrals:  $\int_{-\infty}^{\infty} F(k)e^{-ikx} dk$  and  $\int_{-\infty}^{\infty} G(k')e^{-ik'x} dk'$ .

Parseval's identity is very useful for understanding the physical interpretation of the transform function  $F(k)$  when the physical significance of  $f(x)$  is known, as illustrated in the following example:

*Examples* The displacement of a **damped harmonic oscillator** as a function of time is given by

$$f(t) = \begin{cases} 0 & \text{for } t < 0, \\ e^{-t/\tau} \sin \omega_0 t & \text{for } t \geq 0. \end{cases}$$

The Fourier transform of this function is given by

$$\begin{aligned} F(\omega) &= \int_{-\infty}^0 0 \times e^{-i\omega t} dt + \int_0^{\infty} e^{-t/\tau} \sin \omega_0 t e^{-i\omega t} dt \\ &= 0 + \frac{1}{2i} \int_0^{\infty} \left[ e^{-i(\omega - \omega_0)t - t/\tau} - e^{-i(\omega + \omega_0)t - t/\tau} \right] dt \\ &= \frac{1}{2} \left( \frac{1}{\omega + \omega_0 - i/\tau} - \frac{1}{\omega - \omega_0 - i/\tau} \right). \end{aligned}$$

The physical interpretation of  $|F(\omega)|^2$  is the energy content per unit frequency interval (i.e., the energy spectrum) while  $|f(t)|^2$  is proportional to the sum of the kinetic and potential energies of the oscillator. Hence, Parseval's identity, expressed by

$$\int_{-\infty}^{\infty} |f(t)|^2 dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} |F(\omega)|^2 d\omega,$$

shows the equivalence of these two alternative specifications for the total energy to within a constant.

### 12.1.6 Fourier Transforms in Higher Dimensions

The concept of the Fourier transform can be extended naturally to more than one dimension. For example, in three dimensions we can define the Fourier transform of  $f(x, y, z)$  as

$$F(k_x, k_y, k_z) = \frac{1}{(2\pi)^{3/2}} \iiint f(x, y, z) e^{-ik_x x} e^{-ik_y y} e^{-ik_z z} dx dy dz \quad (12.15)$$

and its inverse as

$$f(x, y, z) = \frac{1}{(2\pi)^{3/2}} \iiint F(k_x, k_y, k_z) e^{ik_x x} e^{ik_y y} e^{ik_z z} dk_x dk_y dk_z. \quad (12.16)$$

Denoting the vector with components  $k_x, k_y, k_z$  by  $\mathbf{k}$  and that with components  $x, y, z$  by  $\mathbf{r}$ , we can write the Fourier transform pair (12.15), (12.16) as follows:

♠ **Fourier transforms in three dimensions:**

$$F(\mathbf{k}) = \frac{1}{(2\pi)^{3/2}} \int f(\mathbf{r}) e^{-i\mathbf{k} \cdot \mathbf{r}} d\mathbf{r},$$

$$f(\mathbf{r}) = \frac{1}{(2\pi)^{3/2}} \int F(\mathbf{k}) e^{i\mathbf{k} \cdot \mathbf{r}} d\mathbf{k}.$$

It is pedagogical to evaluate the Fourier transform of a function  $f(\mathbf{r})$  under the condition that the system possesses spherical symmetry, i.e.,  $f(\mathbf{r}) = f(r)$ . We employ **spherical coordinates** in which the vector  $\mathbf{k}$  of the Fourier transform lies along the polar axis ( $\theta = 0$ ). We then have

$$d\mathbf{r} = r^2 \sin \theta dr d\theta d\phi \quad \text{and} \quad \mathbf{k} \cdot \mathbf{r} = kr \cos \theta,$$

where  $k = |\mathbf{k}|$ . The Fourier transform is then given by

$$\begin{aligned} F(\mathbf{k}) &= \frac{1}{(2\pi)^{3/2}} \int f(\mathbf{r}) e^{-i\mathbf{k} \cdot \mathbf{r}} d\mathbf{r} \\ &= \frac{1}{(2\pi)^{1/2}} \int_0^\infty dr r^2 f(r) \int_0^\pi d\theta \sin \theta e^{-ikr \cos \theta}. \end{aligned}$$

The integral over  $\theta$  may be straightforwardly evaluated by noting that

$$\frac{d}{d\theta} e^{-ikr \cos \theta} = ikr \sin \theta e^{-ikr \cos \theta}.$$

Therefore,

$$\begin{aligned} F(\mathbf{k}) &= \frac{1}{(2\pi)^{1/2}} \int_0^\infty r^2 f(r) dr \left[ \frac{e^{-ikr \cos \theta}}{ikr} \right]_{\theta=0}^{\theta=\pi} \\ &= \frac{1}{(2\pi)^{1/2}} \int_0^\infty 2r^2 f(r) \left( \frac{\sin kr}{kr} \right) dr. \end{aligned}$$

**Remark.** A similar result may be obtained for **two-dimensional Fourier transforms** in which  $f(\mathbf{r}) = f(\rho)$ , i.e.,  $f(\mathbf{r})$  is independent of the azimuthal angle  $\phi$ . In this case, we find

$$F(\mathbf{k}) = \int_0^\infty \rho f(\rho) J_0(k\rho) d\rho,$$

where  $J_0(x)$  is the zeroth order **Bessel function**.

## Exercises

1. Show that if  $f(x)$  is piecewise continuous over  $(a, b)$ , then

$$\lim_{\xi \rightarrow \infty} \int_a^b f(x) \sin \xi x dx = 0.$$

**Solution:** If  $f$  has a continuous derivative, this is easily proved; we integrate by parts to obtain

$$\int_a^b f(x) \cos \xi x dx = \left[ f(x) \frac{\sin \xi x}{\xi} \right]_a^b - \frac{1}{\xi} \int_a^b f'(x) \sin \xi x dx,$$

which tends to zero as  $\xi \rightarrow \infty$  since the integral on the right-hand side is bounded. If  $f$  is not integrable, let  $p$  be a continuously differentiable function such that  $\int_a^b |f(x) - p(x)| dx < \varepsilon$ . Then

$$\begin{aligned} \left| \int_a^b [f(x) - p(x)] \cos \xi x dx \right| &\leq \int_a^b |f(x) - p(x)| |\cos \xi x| dx \\ &\leq \int_a^b |f(x) - p(x)| dx < \varepsilon \end{aligned}$$

independently of  $\xi$ , and as the preceding discussion gave us  $\int_a^b p(x) \cos \xi x dx \rightarrow 0$ , it follows that  $\int_a^b f(x) \cos \xi x dx \rightarrow 0$  as well. The proof that  $\int_a^b p(x) \sin \xi x dx \rightarrow 0$  is similar. ♣

2. Show that  $\int_0^\infty \frac{\sin x}{x} dx = \frac{\pi}{2}$ .

**Solution:** If we substitute  $\lambda = 2\pi$ ,  $x = \pi$  into (11.60) and note that the integrand is an odd function, it follows that

$$\int_0^\pi \frac{\sin\left(\frac{2n+1}{2}u\right)}{\sin(u/2)} du = \pi. \quad (12.17)$$

Applying the result of Exercise 1 noted above to the function  $[(2/u) - 1/\sin(u/2)]$  (which is bounded in  $0 < u < \pi$ ), we have

$$\lim_{n \rightarrow \infty} \int_0^\pi \sin\left(\frac{2n+1}{2}u\right) \left\{ \frac{2}{u} - \frac{1}{\sin(u/2)} \right\} du = 0. \quad (12.18)$$

Summing (12.17) and (12.18), we obtain

$$\lim_{n \rightarrow \infty} \int_0^\pi \frac{2 \sin \frac{2n+1}{2}u}{u} du = \pi.$$

Changing variables and letting  $t = (2n+1)u/2$ , we set

$$\lim_{n \rightarrow \infty} \int_0^{(2n+1)\pi/2} \frac{\sin t}{t} dt = \frac{\pi}{2}.$$

We already know that  $\int_0^M (\sin t)/t dt$  tends to a limit as  $M \rightarrow \infty$  which completes our proof. ♣

3. Show that

$$\lim_{A \rightarrow \infty} \int_0^b f(x) \frac{\sin Ax}{x} dx = \frac{\pi}{2} f(0+) \text{ for } b > 0$$

whenever  $f$  is piecewise smooth.

**Solution:** Observe that

$$\begin{aligned} \int_0^b f(x) \frac{\sin Ax}{x} dx &= \int_0^b f(0+) \frac{\sin Ax}{x} dx + \int_0^b \frac{f(x) - f(0+)}{x} \sin Ax dx \\ &= f(0+) \int_0^{Ab} \frac{\sin u}{u} du + \int_0^b \frac{f(x) - f(0+)}{x} \sin Ax dx. \end{aligned}$$

From the result of Exercise 1, the last integral tends to zero as  $A \rightarrow \infty$ , since the integrand is piecewise smooth in the interval  $0 < x < b$ . It also remains bounded in this interval since, as  $x$  tends to zero,  $[f(x) - f(0+)]/x$  tends to  $f'(0+)$ . From Exercise 2, the other integral tends to the desired value. ♣

## 12.2 Convolution and Correlations

### 12.2.1 Convolution Theorem

In the application of the Fourier transform, we often encounter a product such as  $F(k)G(k)$ , where each of two functions is the Fourier transform of a function  $f(x)$  and  $g(x)$ , respectively. Here, we are interested in finding out how the inverse Fourier transform of the product denoted by

$$\mathcal{F}^{-1}[F(k)G(k)],$$

is related to the individual inverse function

$$\mathcal{F}^{-1}[F(k)] = f(x) \quad \text{and} \quad \mathcal{F}^{-1}[G(k)] = g(x).$$

To begin with, we introduce a key concept called **convolution** and then state an important theorem that plays a central role in the discussion of the matter.

#### ♠ Convolution:

The convolution of the function  $f(x)$  and  $g(x)$ , denoted by  $f * g$ , is defined by

$$f * g = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(u)g(x-u)du. \quad (12.19)$$

The convolution obeys the **commutative**, **associative**, and **distributive** laws of algebra, i.e., if we have function  $f_1, f_2, f_3$ , then

$$\begin{aligned} f_1 * f_2 &= f_2 * f_1 && \text{(Commutative).} \\ f_1 * (f_2 * f_3) &= (f_1 * f_2) * f_3 && \text{(Associative).} \\ f_1 * (f_2 + f_3) &= (f_1 * f_2) + (f_1 * f_3) && \text{(Distributive).} \end{aligned} \quad (12.20)$$

We are now ready to prove the following important theorem regarding the product  $F(k)G(k)$  of two Fourier transforms.

#### ♠ Convolution theorem:

If  $F(k)$  and  $G(k)$  are Fourier transforms of  $f(x)$  and  $g(x)$ , respectively, then

$$F(k)G(k) = \mathcal{F}[f * g]. \quad (12.21)$$

*Proof* It follows from the definition of the Fourier transform that

$$F(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-ikx} dx,$$

$$G(k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(x) e^{-ikx} dx,$$

which yields

$$F(k)G(k) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x)g(x') e^{-ik(x+x')} dx dx'. \quad (12.22)$$

Let  $x+x' = u$  in the double integral of (12.22) transform independent variables from  $(x, x')$  to  $(x, u)$ . We thus have

$$dx dx' = \frac{\partial(x, x')}{\partial(x, u)} du dx,$$

where the Jacobian of the transformation is

$$\frac{\partial(x, x')}{\partial(x, u)} = \begin{vmatrix} \frac{\partial x}{\partial x} & \frac{\partial x}{\partial u} \\ \frac{\partial x'}{\partial x} & \frac{\partial x'}{\partial u} \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} = 1.$$

Then (12.22) becomes

$$\begin{aligned} F(k)G(k) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x)g(u-x) e^{-iku} dx du \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-iku} \left\{ \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x)g(u-x) dx \right\} du \\ &= \mathcal{F}[f * g]. \quad \clubsuit \end{aligned} \quad (12.23)$$

### 12.2.2 Cross-Correlation Functions

There are several important functions related to the convolution, which are called **correlation functions** (see below) and **auto-correlation functions** (see Sect. 12.2.3).

#### ♠ Cross-correlation function:

The cross-correlation of two functions  $f$  and  $g$  is defined by

$$c(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f^*(x)g(x+z)dx. \quad (12.24)$$

Despite the apparent similarity between the cross-correlation function (12.24) and the definition of convolution (12.19), their uses and interpretations are very different: the cross-correlation provides a quantitative measure of the similarity of two functions  $f$  and  $g$  since one is displaced through a distance  $z$  relative to the other.

**Remark.** Similar to the convolution, the cross-correlation is both **associative** and **distributive**. Unlike the convolution, however, it is not **commutative**.

We arrive at an important theorem by considering the Fourier transform of (12.24):

♠ **Wiener–Kinchin theorem:**

The Fourier transform of the cross-correlation of  $f$  and  $g$  is equal to the product of  $F^*(k)$  and  $G(k)$  multiplied by  $\sqrt{2\pi}$ , i.e.,

$$\mathcal{F}[c(z)] \equiv C(k) = F^*(k)G(k). \quad (12.25)$$

*Proof*

$$\begin{aligned} \mathcal{F}[c(x)] \equiv C(k) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dz e^{-ikz} \left\{ \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f^*(x)g(z+x)dx \right\} \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dx f^*(x) \left\{ \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(z+x)e^{-ikz} dz \right\}. \end{aligned}$$

Making the substitution  $u = z + x$  in the second integral, we obtain

$$\begin{aligned} C(k) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dx f^*(x) \left\{ \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(u)e^{-ik(u-x)} du \right\} \\ &= \left\{ \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f^*(x)e^{ikx} dx \right\} \left\{ \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(u)e^{-iku} du \right\} \\ &= F^*(k)G(k). \quad \clubsuit \end{aligned} \quad (12.26)$$

It readily follows from the definition (12.24) and the theorem (12.25) that

$$c(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} C(k)e^{ikz} dk = \int_{-\infty}^{\infty} F^*(k)G(k)e^{ikz} dk. \quad (12.27)$$

Then, setting  $z = 0$  gives us the multiplication theorem

$$\int_{-\infty}^{\infty} f^*(x)g(x)dx = \int_{-\infty}^{\infty} F^*(k)G(k)dk. \quad (12.28)$$

Further, by letting  $g = f$ , we arrive at the following identity:

♠ **Plancherel identity:**

A function  $f(x)$  and its Fourier transform  $F(k)$  are related to one another by the identity

$$\int_{-\infty}^{\infty} |f(x)|^2 dx = \int_{-\infty}^{\infty} |F(k)|^2 dk, \quad (12.29)$$

which is called the **Plancherel identity**.

Plancherel's identity is sometimes called Parseval's identity, aims to the analogy with Fourier series.

### 12.2.3 Autocorrelation Functions

Particularly when  $g(x) = f(x)$ , the cross-correlation function  $c(z)$  is referred to specifically as follows:

♠ **Autocorrelation function:**

The autocorrelation function of  $f(x)$  is defined by

$$a(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f^*(x) f(x+z) dx.$$

Using the Wiener–Kinchin theorem (12.26), we see that

$$\begin{aligned} a(z) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} A(k) e^{ikx} dk = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \sqrt{2\pi} F^*(k) F(k) e^{ikx} dk \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |F(k)|^2 e^{ikx} dk. \end{aligned}$$

This implies that the quantity  $|F(k)|^2$ , called the **power spectrum** of  $f(x)$ , is the Fourier transform of the autocorrelation function as formally stated below.

♠ **Power spectrum:**

Given  $f(x)$ , we have

$$|F(k)|^2 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} a(z) e^{-ikx} dz,$$

where  $F(k)$  and  $a(z)$  are, respectively, the Fourier transform and the autocorrelation function of  $f(x)$ .

This result is frequently made use of in practical applications of Fourier transforms.

## 12.3 Discrete Fourier Transform

### 12.3.1 Definitions

The present section includes several topics associated with numerical computation of Fourier transforms. Generally, in computational work, we do not treat a continuous function  $f(t)$ , but rather  $f(t_n)$  given by a discrete set of  $t_n$ 's. (For now, we assume that a physical process of interest is described in the time domain.) In most common situations, the value of  $f(t)$  is recorded at evenly spaced intervals. In this context, we have to estimate the Fourier transform of a function from a finite number of its sampled points.

Suppose that we have a set of measurements performed at equal time intervals of  $\Delta$ . Then the sequence of sampled values is given by

$$f_k = f(t_k), \quad t_k = k\Delta \quad (k = 0, 1, 2, \dots, N-1). \quad (12.30)$$

For simplicity, we assume that  $N$  is even. With  $N$  numbers of input, we can produce at most  $N$  independent numbers of output. So, instead of trying to estimate the Fourier transform  $F(\omega)$  in the whole range of frequency  $\omega$ , we seek estimates only at the discrete values  $\omega = \omega_n$  with  $n = 0, 1, \dots, N-1$ . By analogy with the Fourier transform for a continuous function  $f(t)$ , we may define the Fourier transform for a discrete set of  $f_k = f(t_k)$  ( $k = 0, 1, \dots, N-1$ ) as below.

#### ♠ Discrete Fourier transform:

The discrete Fourier transform for a discrete set of  $f_k$  given by (12.30) is defined by

$$F_n = F(\omega_n) = \frac{1}{N} \sum_{k=0}^{N-1} f(t_k) e^{-i\omega_n t_k} = \frac{1}{N} \sum_{k=0}^{N-1} f_k e^{-2\pi i k n / N}, \quad (12.31)$$

with the definition

$$\omega_n \equiv \frac{2\pi n}{N\Delta} \quad (n = 0, 1, \dots, N-1). \quad (12.32)$$

Note that  $F_n$  is associated with frequency  $\omega_n$ . Of importance is the fact that in (12.31),  $n$  can be any integer from  $-\infty$  to  $\infty$ , whereas  $k$  in (12.31) runs

from 0 to  $N - 1$ . The latter restriction is due to the fact that  $F_n$  is periodic with a period of  $N$  terms. In fact, for any integer  $n$  such that  $0 \leq n \leq N - 1$ , we have

$$F_n = F_{n \pm N} = F_{n \pm 2N} = \cdots$$

as readily follows from (12.31).

### 12.3.2 Inverse Transform

Given the discrete transform  $F_n$ , we can reproduce the time series  $f_k$  with the aid of the inverse relationship:

#### ♠ Inverse of discrete Fourier transform:

The discrete Fourier transform of a set  $\{f_k\}$  satisfies the relation

$$f_k = \sum_{n=0}^{N-1} F_n e^{2\pi i k n / N}. \quad (12.33)$$

*Proof* For the proof, it suffices to observe that

$$\sum_{n=0}^{N-1} e^{-2\pi i n(k-k')/N} = \begin{cases} N & (k = k'), \\ 0 & (\text{otherwise}). \end{cases} \quad (12.34)$$

(see Exercise 1 in Sect. 12.3). Then, from (12.31) and (12.34), we have

$$\begin{aligned} \sum_{n=0}^{N-1} F_n e^{2\pi i n k' / N} &= \frac{1}{N} \sum_{n=0}^{N-1} \sum_{k=0}^{N-1} f_k e^{-2\pi i n(k-k')/N} \\ &= \frac{1}{N} \sum_{k=0}^{N-1} f_k \cdot N \delta_{kk'} = f_{k'}. \quad \clubsuit \end{aligned}$$

Note that the only differences between expressions (12.31) and (12.33) for  $F_n$  and  $f_k$ , respectively, are (i) changing the sign in the exponential, and (ii) dividing the answer by  $N$ . This means that a computational procedure for calculating discrete Fourier transforms can, with slight modifications, also be used to calculate the inverse transform. In addition, we see from the inverse transform that only  $N$  values of the frequency  $\omega_n$  are needed and that they range from 0 to  $N - 1$ , just as with the discrete time  $t_k$ .

### 12.3.3 Nyquist Frequency and Aliasing

In the above discussion, we have taken the view that the index  $n$  in (12.31) varies from 0 to  $N$ . In this convention,  $n$  in  $F_n$  and  $k$  in  $f_k$  vary over exactly the same range, so the mapping of  $N$  numbers into  $N$  numbers is manifest. Alternatively, since the quantity  $F_n$  given in (12.31) is periodic in  $n$  with period  $N$  (i.e.,  $F_n = F_{N+n}$ ),  $n$  in  $F_n$  is allowed to vary from  $-N/2$  to  $(N/2) - 1$ . In the latter convention, the discrete Fourier transform and its inverse transform read, respectively,

$$F_n = \sum_{k=-N/2}^{N/2-1} f_k e^{-2\pi i k n / N} \quad \text{and} \quad f_k = \frac{1}{N} \sum_{n=-N/2}^{N/2-1} F_n e^{2\pi i k n / N}. \quad (12.35)$$

Emphasis is placed on the fact that in (12.35), the upper bound of the summation is not  $N/2$  but  $(N/2) - 1$ . This ensures the count of  $\omega_n$  to  $N$ . Indeed, the periodicity of  $F_n$  in  $n$  with the period  $N$  implies that the descretized frequency  $\omega_n = 2\pi n / (N\Delta)$  is also periodic in  $n$  with  $N$ . Hence, the two extreme values of  $\omega_n$ , i.e.,

$$\omega_{-N/2} = -\frac{\pi}{\Delta} \quad \text{and} \quad \omega_{N/2} = \frac{\pi}{\Delta},$$

contribute to  $F_n$  as given in (12.31) in the same way. These two indistinguishable frequencies are known as the **Nyquist critical frequencies**.

♠ **Nyquist critical frequency:**

A Nyquist critical frequency is defined by

$$\omega_c \equiv \frac{\pi}{\Delta},$$

where  $\Delta$  is the sampling interval:  $t_k = k\Delta$  ( $k = 0, 1, \dots, N-1$ ).

The Nyquist critical frequency has the following peculiarity. Suppose that we sample a sine wave of the Nyquist critical frequency, expressed by

$$f(t) = \sin(\omega_c t),$$

at the sampling interval  $\Delta$ . Then we have

$$f_k = f(t_k) = \sin(\omega_c t_k + \theta) = \sin\left[\frac{\pi}{\Delta}(k\Delta + \theta)\right] = \sin(k\pi + \theta) \\ (k = 0, 1, \dots, N-1),$$

where  $\theta$  is determined by the initial condition:  $f(0) = \sin \theta$ . Then, the sampling becomes two sample points per cycle:  $\sin \theta$  and  $-\sin \theta$ .

The above arguments further suggest that descretized frequencies  $\omega_n$  above (and below)  $\omega_c$  are identified with  $\omega_{n-N}$  (and  $\omega_{n+N}$ ). This phenomenon, peculiar to discrete sampling, leads to the following important consequence:

♠ **Aliasing:**

When a continuous function  $f(t)$  is sampled with an interval  $\Delta$ , all of the power spectral density lying outside of the range  $[-\omega_c, \omega_c)$  with  $\omega_c = \pi/\Delta$  is moved into that range. Owing to a phenomenon called **aliasing**.

Through discrete sampling, therefore, any frequency component outside of the range  $[-\omega_c, \omega_c)$  is falsely translated into that range.

*Example* Suppose that two continuous waves  $\exp(i\omega_1 t)$  and  $\exp(i\omega_2 t)$  are sampled with the same interval  $\Delta$ . Then, if  $\omega_2 = \omega_1 \pm 2\omega_c$ , we obtain the same samples, since

$$\begin{aligned}\exp(i\omega_2 t_k) &= \exp(i\omega_1 t_k) \times \exp(\pm 2i\omega_c t_k) \\ &= \exp(i\omega_1 t_k) \times \exp(\pm 2k\pi i) = \exp(i\omega_1 t_k),\end{aligned}$$

where  $t_k = k\Delta$  ( $k = 0, 1, \dots, N-1$ ). Hence, a sinusoidal wave having a frequency lying outside the range  $[-\omega_c, \omega_c)$  appears the same as the sinusoidal wave whose frequency is within the range.

*Remark.* The way to overcome aliasing is to (i) know the natural bandwidth limit of the signal – or else enforce a known limit by analog filtering of the continuous signal, and then (ii) sample at a rate sufficiently rapid to give at least two points per cycle of the highest frequency present.

### 12.3.4 Sampling Theorem

We present below a famous theorem that is useful in certain applications of the discrete Fourier transform.

♠ **Sampling theorem:**

Suppose that a continuous function  $f(t)$  is sampled at an interval  $\Delta$  as  $f_k = f(k\Delta)$ . If its Fourier transform satisfies the condition that  $F(\omega) = 0$  for all  $|\omega| \geq \omega_c = \pi/\Delta$ , then we have

$$f(t) = \sum_{k=-\infty}^{\infty} f_k \frac{\sin[\omega_c(t - k\Delta)]}{\pi(t - k\Delta)}.$$

This theorem states that if a signal  $f(t)$  that is in question is bandwidth-limited (i.e.,  $F(\omega) = 0$  for  $|\omega| \geq |\omega_0|$ ) with a certain preassigned frequency  $\omega_0$ , then the entire information content of the signal can be recorded by sampling it at the interval  $\Delta = \pi/\omega_0$ .

*Proof* Given a continuous function  $f(t)$ , we express it by the inverse Fourier transform as

$$f(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} F(\omega) e^{i\omega t} d\omega.$$

From hypothesis,  $F(\omega)$  vanishes at  $\omega \geq |\omega_c|$  so that

$$f(t) = \frac{1}{\sqrt{2\pi}} \int_{-\omega_c}^{\omega_c} F(\omega) e^{i\omega t} d\omega,$$

which yields

$$f(t_k) = \frac{1}{\sqrt{2\pi}} \int_{-\omega_c}^{\omega_c} F(\omega) e^{i\omega t_k} d\omega \quad \text{for } t_k = k\Delta \quad (k \in \mathbf{Z}).$$

Consider the Fourier series expansion of  $F(\omega)$  as

$$F(\omega) = \sum_{k=-\infty}^{\infty} c_k e^{-i\omega t_k} \quad \text{for } |\omega| \leq \omega_c, \quad (12.36)$$

where the coefficients  $c_k$  read

$$c_k = \frac{1}{\sqrt{2\pi}} \int_{-\omega_c}^{\omega_c} F(\omega) e^{i\omega t_k} d\omega = f(t_k). \quad (12.37)$$

From (12.36) and (12.37), we obtain

$$F(\omega) = \sum_{k=-\infty}^{\infty} f(t_k) e^{-i\omega t_k} \quad \text{for } |\omega| \leq \omega_c.$$

Now we define

$$H(\omega) = \sum_{k=-\infty}^{\infty} f(t_k) e^{-i\omega t_k} \quad \text{for all } \omega.$$

While the function  $H(\omega)$  is a periodic function with period  $2\omega_c$ , the  $F(\omega)$  is identically zero outside the interval  $[-\omega_c, \omega_c]$ . This being so, we can write

$$F(\omega) = H(\omega)S(\omega) \quad \text{with } S(\omega) = \begin{cases} 1 & |\omega| \leq \omega_c, \\ 0 & |\omega| > \omega_c. \end{cases}$$

Thus we have

$$F(\omega) = \sum_{k=-\infty}^{\infty} f(t_k) e^{-i\omega t_k} S(\omega),$$

and its inverse transform reads

$$\begin{aligned}
 f(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left[ \sum_{k=-\infty}^{\infty} f(t_k) e^{-i\omega t_k} S(\omega) \right] e^{i\omega t} d\omega \\
 &= \sum_{k=-\infty}^{\infty} f(t_k) \frac{1}{\sqrt{2\pi}} \int_{-\omega_c}^{\omega_c} e^{i\omega(t-t_k)} d\omega \\
 &= \sum_{k=-\infty}^{\infty} f(t_k) \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} S(\omega) e^{i\omega(t-t_k)} d\omega \\
 &= \sum_{k=-\infty}^{\infty} f(t_k) \frac{\sin[\omega_c(t-t_k)]}{\omega_c(t-t_k)}. \quad \clubsuit
 \end{aligned}$$

### 12.3.5 Fast Fourier Transform

The **fast Fourier transform** (often abbreviated by **FFT**) is an algorithm for calculating discrete Fourier transforms and is widely known as a useful tool in computational physics. In this subsection, we demonstrate the efficiency of this computational method.

In a typical discrete Fourier transform, one has a sum of  $N$  terms expressed by

$$F_n = \sum_{k=0}^{N-1} W^{nk} f_k, \quad (12.38)$$

where  $W$  is a complex number defined by

$$W \equiv e^{2\pi i/N}.$$

Notably, the left-hand side of (12.38) can be regarded as a product of the vector consisting of the elements  $\{f_k\}$  with a matrix whose  $(n, k)$ th element is the constant  $W$  to the power  $n \times k$ . The matrix multiplication produces a vector whose components are the  $F_n$ 's. This operation evidently requires  $N^2$  complex-number multiplications plus a smaller number of operations to generate the required powers of  $W$ . Thus, the discrete Fourier transform appears to be an  $O(N^2)$  process.

The efficiency of the fast Fourier transform manifests in the fact that it enables us to compare the discrete Fourier transform in  $O(N \log_2 N)$  operations. The difference between  $O(N^2)$  and  $O(N \log_2 N)$  is immense. With  $N = 10^8$ , e.g., it is the difference between, roughly, 2 s and 3 months of CPU time on a gigahertz cycle computer.

The fast Fourier transform is based on the fact that a discrete Fourier transform of length  $N$  can be rewritten as the sum of two discrete Fourier transforms, each of length  $N/2$ . This is easily seen from (12.38) as follows:

$$\begin{aligned}
F_n &= \sum_{k=0}^{N-1} e^{2\pi i k n / N} f_k \\
&= \sum_{k=0}^{N/2-1} e^{2\pi i n (2k) / N} f_{2k} + \sum_{k=0}^{N/2-1} e^{2\pi i n (2k+1) / N} f_{2k+1} \\
&= \sum_{k=0}^{N/2-1} e^{2\pi i n k / (N/2)} f_{2k} + W^n \sum_{k=0}^{N/2-1} e^{2\pi i n k / (N/2)} f_{2k+1} \\
&= F_n^e + W^n F_n^o.
\end{aligned} \tag{12.39}$$

Here  $W$  is the same complex constant we defined in (12.38). The  $F_n^e$  denotes the  $n$ th component of the Fourier transform of the sequence  $(f_{2k})$  with length  $N/2$  expressed by

$$(f_{2k}) = (f_0, f_2, f_4, \dots, f_{N-2}),$$

which consists of even components of the original  $f_k$ 's. Similarly, the  $F_n^o$  is the corresponding transform of length  $N/2$  formed from odd components. Recall that  $F_n$  is periodic in  $n$  with the period  $N$ . On the other hand, the transforms  $F_n^e$  and  $F_n^o$  are periodic in  $k$  with length  $N/2$ . This period-reduction property is the origin of the efficiency of the fast Fourier transform as demonstrated below.

Having decomposed  $F_n$  into  $F_n^e$  and  $F_n^o$ , we can apply the same procedure to  $F_n^e$  and  $F_n^o$  to produce  $N/4$  even-numbered and odd-numbered data:

$$\begin{aligned}
F_n^e &= \sum_{k=0}^{N/4-1} e^{2\pi i n k / (N/4)} f_{4k} + W^n \sum_{k=0}^{N/4-1} e^{2\pi i n k / (N/4)} f_{4k+2} \\
&= F_n^{ee} + W^n F_n^{eo},
\end{aligned} \tag{12.40}$$

$$\begin{aligned}
F_n^o &= \sum_{k=0}^{N/4-1} e^{2\pi i n k / (N/4)} f_{4k+1} + W^n \sum_{k=0}^{N/4-1} e^{2\pi i n k / (N/4)} f_{4k+3} \\
&= F_n^{oe} + W^n F_n^{oo}.
\end{aligned} \tag{12.41}$$

Here, the  $F_k^{eo}$ , e.g., is the transform of the sequence  $(f_{4k+2})$  given by

$$(f_{4k+2}) = (f_2, f_6, \dots, f_{N-2}),$$

whose length is  $N/4$ . We can continue the above procedure until we obtain the transform of a single-point sequence, say,

$$F_n^{eoeoeoe\cdots oee} = f_k \quad \text{for some } k. \tag{12.42}$$

This implies that for every pattern of  $\log_2 N$   $e$ 's and  $o$ 's, there is a one-point transform that is just one of the input numbers  $f_k$ . Therefore, by relating all the terms  $f_k$  ( $0 \leq k \leq N-1$ ) to  $\log_2 N$  patterns of  $e$ 's and  $o$ 's and

then tracking back to the procedures (12.39), (12.40), (12.41), and (12.42) to reproduce  $F_n$ , we will successfully obtain the discrete Fourier transform  $F_n$  ( $0 \leq n \leq N-1$ ) of the original data  $f_k$  ( $0 \leq k \leq N-1$ ).

One may ask a question as to the way we can figure out which value of  $k$  corresponds to which pattern of  $e$ 's and  $o$ 's in (12.42). As we demonstrate later, this can be achieved by reversing the pattern of  $e$ 's and  $o$ 's and setting  $e = 0$  and  $o = 1$ . Then, we have the corresponding value of  $k$  in a binary expression. This idea of **bit reversal** can be exploited in a very clever way that makes FFTs practical.

### 12.3.6 Matrix Representation of FFT Algorithm

To make our discussion more concrete, we now present an actual FFT procedure to obtain the discrete Fourier transform  $F(n)$  ( $n = 0, 1, 2, 3$ ) of the original vector data  $f(k)$  ( $k = 0, 1, 2, 3$ ). By definition,  $F(n)$  is given in the matrix representation as

$$\begin{bmatrix} F(0) \\ F(1) \\ F(2) \\ F(3) \end{bmatrix} = \begin{bmatrix} W^0 & W^0 & W^0 & W^0 \\ W^0 & W^1 & W^2 & W^3 \\ W^0 & W^2 & W^4 & W^6 \\ W^0 & W^3 & W^6 & W^9 \end{bmatrix} \begin{bmatrix} f(0) \\ f(1) \\ f(2) \\ f(3) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & W^1 & W^2 & W^3 \\ 1 & W^2 & W^0 & W^2 \\ 1 & W^3 & W^2 & W^1 \end{bmatrix} \begin{bmatrix} f(0) \\ f(1) \\ f(2) \\ f(3) \end{bmatrix}, \quad (12.43)$$

where we used the fact that

$$W^4 = \left(e^{2\pi i/4}\right)^4 = e^{2\pi i} = 1;$$

More generally, we have

$$W^{nk} = W^{nk \bmod(N)},$$

where the number

$$nk \bmod(N)$$

is the remainder when the integer  $nk$  is divided by  $N$ . The trick involved in the FFT algorithm is to decompose the product of the vector and the matrix appearing in (12.43) into that of a vector and two matrices:

$$\begin{bmatrix} F(0) \\ F(2) \\ F(1) \\ F(3) \end{bmatrix} = \begin{bmatrix} 1 & W^0 & 0 & 0 \\ 1 & W^2 & 0 & 0 \\ 0 & 0 & 1 & W^1 \\ 0 & 0 & 1 & W^3 \end{bmatrix} \begin{bmatrix} 1 & 0 & W^0 & 0 \\ 0 & 1 & 0 & W^0 \\ 1 & 0 & W^2 & 0 \\ 0 & 1 & 0 & W^1 \end{bmatrix} \begin{bmatrix} f(0) \\ f(1) \\ f(2) \\ f(3) \end{bmatrix}. \quad (12.44)$$

The equivalence between (12.43) and (12.44) is verified in a straightforward manner. Nevertheless, the reader should pay attention to the fact that in (12.44), the order of elements in the vector  $F(n)$  is altered from that in the

original form (12.43). As we demonstrate later, this altering property of the order of  $F(n)$  enables us to compute efficiently the  $F(n)$  from  $f(k)$  with the help of the **bit-reversing process**.

The efficiency of FFT can be observed by counting up the number of multiplication (and additions) between matrix elements in order to complete the matrix operation given in (12.44). First we set

$$\begin{bmatrix} f_1(0) \\ f_1(1) \\ f_1(2) \\ f_1(3) \end{bmatrix} = \begin{bmatrix} 1 & 0 & W^0 & 0 \\ 0 & 1 & 0 & W^0 \\ 1 & 0 & W^2 & 0 \\ 0 & 1 & 0 & W^1 \end{bmatrix} \begin{bmatrix} f_0(0) \\ f_0(1) \\ f_0(2) \\ f_0(3) \end{bmatrix},$$

in which  $f_0(k) = f(k)$  ( $k = 0, 1, 2, 3$ ). Then  $f_1(0)$  is obtained through one complex-number multiplication and one complex-number addition, i.e.,

$$f_1(0) = f_0(0) + W^0 f_0(2). \quad (12.45)$$

We can obtain  $f_1(1)$  in the same manner as above. On the contrary, to obtain  $f_1(2)$ , only one complex-number addition is needed due to the relation  $W^2 = -W^0$ . In fact,

$$f_1(2) = f_0(0) + W^2 f_0(2) = f_0(0) - W^0 f_0(2),$$

in which the product  $W^0 f_0(2)$  was evaluated earlier in the calculation of (12.45). Likewise,  $f_1(3)$  is also computed by only one addition owing to the relation  $W^3 = -W^1$ . As a consequence, the vector  $f_1(k)$  ( $k = 0, 1, 2, 3$ ) is calculated through four-times additions and two-times multiplications.

A similar scenario can apply to the remaining computation:

$$\begin{bmatrix} F(0) \\ F(2) \\ F(1) \\ F(3) \end{bmatrix} = \begin{bmatrix} f_2(0) \\ f_2(1) \\ f_2(2) \\ f_2(3) \end{bmatrix} = \begin{bmatrix} 1 & W^0 & 0 & 0 \\ 1 & W^2 & 0 & 0 \\ 0 & 0 & 1 & W^1 \\ 0 & 0 & 1 & W^3 \end{bmatrix} \begin{bmatrix} f_1(0) \\ f_1(1) \\ f_1(2) \\ f_1(3) \end{bmatrix}.$$

Calculation of each number  $f_2(0)$  and  $f_2(2)$  requires both one addition and one multiplication, whereas for  $f_2(1)$  and  $f_2(3)$  only one addition is required because of the relations  $W^2 = -W^0$  and  $W^3 = -W^1$ . Therefore, the entire computation to yield  $F(n)$  in the above context requires *four-time multiplications* and *eight-time additions*. This computational cost is significantly small compared with the direct matrix calculation given in (12.43), where *16-times multiplications* and *12-times additions* are needed. More generally, when considering the transform  $F(k)$  of the length  $N = 2^\gamma$ , the FFT procedure requires the multiplications of  $N\gamma/2$  times and the additions of  $N\gamma$  times, whereas the direct matrix calculation procedure demands  $N^2$ -times multiplications and  $N(N-1)$ -times additions. Thus the superiority of FFT method is considerably enhanced when  $N \gg 1$ .

### 12.3.7 Decomposition Method for FFT

It is still unclear as to how we can find an appropriate decomposition of general  $N \times N$  matrices as performed in (12.44). To see this, we express the indices  $n$  and  $k$  in terms of two-digit expressions:

$$n = 2n_1 + n_0, \quad k = 2k_1 + k_0,$$

where each  $n_1, n_0, k_1, k_0$  takes the value 0 or 1 [e.g.,  $n = 3$  corresponds to  $(n_0, n_1) = (1, 0)$ ]. Then, the discrete Fourier transform reads

$$F(n_1, n_0) = \sum_{k_0=0}^1 \sum_{k_1=0}^1 f_0(k_1, k_0) W^{(2n_1+n_0)(2k_1+k_0)}.$$

Now we apply the identity

$$W^{(2n_1+n_0)(2k_1+k_0)} = W^{4n_1k_1} W^{2n_0k_1} W^{(2n_1+n_0)k_0} = W^{2n_0k_1} W^{(2n_1+n_0)k_0}$$

to obtain

$$F(n_1, n_0) = \sum_{k_0=0}^1 \left[ \sum_{k_1=0}^1 f_0(k_1, k_0) W^{2n_0k_1} \right] W^{(2n_1+n_0)k_0}. \quad (12.46)$$

Denoting the sum in the square bracket by  $f_1(n_0, k_0)$ , we have

$$f_1(n_0, k_0) = \sum_{k_1=0}^1 f_0(k_1, k_0) W^{2n_0k_1}, \quad (12.47)$$

or equivalently,

$$\begin{aligned} f_1(0, 0) &= f_0(0, 0) + f_0(1, 0)W^0, \\ f_1(0, 1) &= f_0(0, 1) + f_0(1, 1)W^0, \\ f_1(1, 0) &= f_0(0, 0) + f_0(1, 0)W^2, \\ f_1(1, 1) &= f_0(0, 1) + f_0(1, 1)W^2. \end{aligned}$$

This system of equations is expressed in matrix form as

$$\begin{bmatrix} f_1(0, 0) \\ f_1(0, 1) \\ f_1(1, 0) \\ f_1(1, 1) \end{bmatrix} = \begin{bmatrix} 1 & 0 & W^0 & 0 \\ 0 & 1 & 0 & W^0 \\ 1 & 0 & W^2 & 0 \\ 0 & 1 & 0 & W^1 \end{bmatrix} \begin{bmatrix} f_0(0, 0) \\ f_0(0, 1) \\ f_0(1, 0) \\ f_0(1, 1) \end{bmatrix}.$$

Similarly, from (12.46) and (12.47), it follows that

$$\begin{bmatrix} f_2(0,0) \\ f_2(0,1) \\ f_2(1,0) \\ f_2(1,1) \end{bmatrix} = \begin{bmatrix} 1 & W^0 & 0 & 0 \\ 1 & W^2 & 0 & 0 \\ 0 & 0 & 1 & W^1 \\ 0 & 0 & 1 & W^3 \end{bmatrix} \begin{bmatrix} f_1(0,0) \\ f_1(0,1) \\ f_1(1,0) \\ f_1(1,1) \end{bmatrix}.$$

Hence, we have

$$F(n_1, n_0) = f_2(n_0, n_1),$$

in which the order of  $n_0$  and  $n_1$  in the parentheses differs on the two sides. This indicates that the individual numbers  $f_2(n_0, n_1)$  are in order not of  $n = 2n_1 + n_0$ , but of the numbers obtained by bit-reversing  $n$ , which is why the **bit-reversing process** is required to obtain the discrete Fourier transform  $F(n)$  using FFT. The above discussion also clearly demonstrates the way to construct the decomposed product of matrices that makes the entire computations a fast.

## Exercises

1. Show that

$$\sum_{n=0}^{N-1} e^{-2\pi i n(k-k')/N} = \begin{cases} N & \text{if } k = k', \\ 0 & \text{otherwise,} \end{cases}$$

where  $k$  and  $k'$  are integers ranging from 0 to  $N - 1$ .

**Solution:** The proof for the case of  $k = k'$  is trivial. When  $k \neq k'$ , then

$$e^{-2\pi i n(k-k')/N} \neq 1 \quad \text{and} \quad e^{-2\pi i n(k-k')} = 1$$

for any choice of  $k$  and  $k'$ , so that we have

$$\sum_{n=0}^{N-1} e^{-2\pi i n(k-k')/N} = \frac{1 - e^{-2\pi i N(k-k')/N}}{1 - e^{-2\pi i (k-k')/N}} = 0. \quad \clubsuit$$

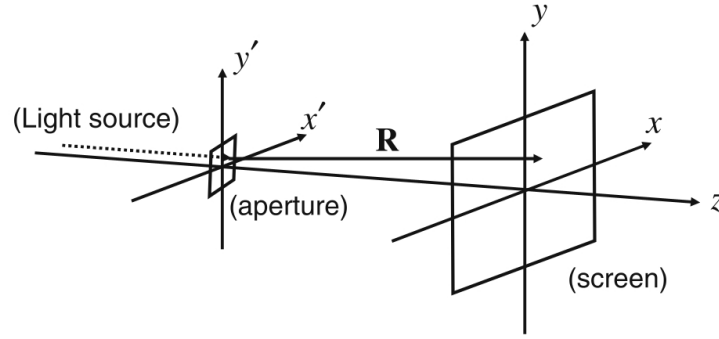
## 12.4 Applications in Physics and Engineering

### 12.4.1 Fraunhofer Diffraction I

In optics, Fourier transformation is a powerful tool to describe an important class of wave diffractions, called **Fraunhofer diffraction**; this refers to the diffraction of electromagnetic radiation observed at a point far from a slit or an aperture. A Fraunhofer diffraction pattern can be described by using the

wave theory of light, which predicts the areas of constructive and destructive interference.

Let us derive the diffraction pattern produced by a rectangular aperture with width  $a$  and height  $b$ . We assume that both incident and diffracted waves can be approximated as being plain waves with wavelength  $\lambda$ . In order to make this assumption, the diffracting obstacle and the observation point must be sufficiently far from the light source so that the curvature of the incident and diffracted light can be neglected (see Fig. 12.1). According to elementary wave optics, the amplitude of light at  $\mathbf{R}$  on the screen is given by



**Fig. 12.1.** Configurations of the light source, a rectangular aperture, and the screen

$$u(\mathbf{R}) = -\frac{ik}{2\pi R} \int_{\Delta S'} u(\mathbf{r}') e^{ik|\mathbf{R}-\mathbf{r}'|} d\mathbf{r}'.$$

Here,  $k = 2\pi/\lambda$ ,  $\Delta S'$  represents the area of the rectangular aperture through which light passes and  $u(\mathbf{r}')$  is the amplitude of the incident wave at  $\mathbf{r}'$  within the aperture:

$$u(\mathbf{r}') = Ae^{i\mathbf{k} \cdot \mathbf{r}'}.$$

We assume that this incident wave is oriented in the direction of the  $z$ -axis. Then, the wave vector  $\mathbf{k}$  is perpendicular to the position vector  $\mathbf{r}'$  so that

$$u(\mathbf{r}') = A = \text{const.}$$

Hence, we have

$$u(\mathbf{R}) = -\frac{ikA}{2\pi R} \int_{-a}^a dx' \int_{-b}^b dy' e^{ik|\mathbf{R}-\mathbf{r}'|}. \quad (12.48)$$

Set  $\mathbf{R} = (x, y, z)$  and  $\mathbf{r}' = (x', y', 0)$ , where the origin is located at the center of the aperture. Under the assumption that  $z \gg |x|, |y|$  and  $|x|, |y| \gg |x'|, |y'|$ , we have

$$\begin{aligned}
|\mathbf{R} - \mathbf{r}'| &= \sqrt{z^2 + (x - x')^2 + (y - y')^2} \\
&= \sqrt{R^2 - 2(xx' + yy') + x'^2 + y'^2} \\
&\simeq R \left( 1 - \frac{xx' + yy'}{R^2} + \frac{x'^2 + y'^2}{2R^2} \right) \simeq R \left( 1 - \frac{xx' + yy'}{R^2} \right).
\end{aligned}$$

Substituting this into (12.48) yields

$$\begin{aligned}
u(\mathbf{R}) &= -\frac{ikA}{2\pi R} e^{ikR} \int_{-a}^a e^{-\frac{ikx}{R}x'} dx' \int_{-b}^b e^{-\frac{iky}{R}y'} dy' \\
&= -\frac{2ikA}{\pi R} e^{ikR} \frac{\sin \frac{kax}{R}}{\frac{kx}{R}} \frac{\sin \frac{kby}{R}}{\frac{ky}{R}}.
\end{aligned}$$

The light intensity distribution  $I(\mathbf{R})$  on the screen is thus given by

$$I(\mathbf{R}) \equiv |u(\mathbf{R})|^2 \propto k^2 a^2 b^2 \left[ \frac{\sin(ka\tilde{x})}{ka\tilde{x}} \frac{\sin(kb\tilde{y})}{kb\tilde{y}} \right]^2,$$

where

$$\tilde{x} = \frac{x}{R}, \quad \tilde{y} = \frac{y}{R}.$$

Remember that  $(\sin \xi)/\xi = 0$  at  $\xi = \pm n\pi$  with integers  $n = 1, 2, \dots$ . In addition, since  $k = 2\pi/\lambda$ , we conclude that

$$I(\mathbf{R}) = 0 \quad \text{at} \quad \tilde{x} = \pm \frac{m\lambda}{2a} \quad \text{or} \quad \tilde{y} = \pm \frac{n\lambda}{2b}, \quad (m, n = 1, 2, \dots),$$

which describes the diffraction pattern generated on the screen.

### 12.4.2 Fraunhofer Diffraction II

We next consider the case of a circular aperture with radius  $a$ . For convenience, we use the polar coordinates defined by  $x = r \cos \theta$ ,  $y = r \sin \theta$ . Then (12.49) reads

$$\begin{aligned}
u(\mathbf{R}) &\propto e^{ikR} \int_{\Delta S'} \exp \left[ \frac{-ik(xx' + yy')}{R} \right] d\mathbf{r}' \\
&= \int_0^a dr' \int_0^{2\pi} d\theta' r' \exp \left[ \frac{-ikrr'(\cos \theta \cos \theta' + \sin \theta \sin \theta')}{R} \right] \\
&= \int_0^a dr' \int_0^{2\pi} d\theta' r' \exp \left[ \frac{-ikrr' \cos(\theta' - \theta)}{R} \right].
\end{aligned}$$

To make it concise, we use the following formulae based on the **Bessel function**  $J_n(x)$ :

$$\int_0^{2\pi} e^{i\zeta \cos \phi} d\phi = 2\pi J_0(\zeta), \quad \int_0^\eta \zeta J_0(\zeta) d\zeta = \eta J_1(\eta).$$

These give us

$$u(\mathbf{R}) \propto 2\pi a^2 \frac{J_1\left(\frac{kar}{R}\right)}{\frac{kar}{R}},$$

where the explicit form of  $J_1(x)$  is obtained from the definition of  $J_\nu(x)$ ,

$$J_\nu(x) = \left(\frac{x}{2}\right)^\nu \sum_{\ell=0}^{\infty} \frac{(-1)^\ell (x/2)^{2\ell}}{\ell! \Gamma(\nu + \ell + 1)},$$

and thus  $\lim_{x \rightarrow 0} J_1(x)/x = 1/2$ . The first zero of  $J_1(x)$  is located at  $x \simeq 1.22\pi$ . Therefore, the radius  $r_0$  of the innermost dark ring on the screen is given by

$$\frac{kar_0}{R} \simeq 1.22\pi, \quad \text{i.e.,} \quad r_0 \simeq \frac{0.61\lambda}{a} R.$$

### 12.4.3 Amplitude Modulation Technique

We conclude this chapter with a discussion regarding the use of Fourier transformations in an **amplitude modulation** (AM) technique. This technique is used in electronic communication, most commonly for transmitting information via a radio carrier wave. As the name indicates, AM works by modulating the vibrational amplitude of the transmitted signal according to the information being sent. This is in contrast to the **frequency modulation** (FM) technique that is also commonly used for transmitting sound, but by modulating its frequency.

For AM, we use two kinds of waves: a **carrier wave**  $c(t)$  and a **message wave**  $m(t)$  that contains information on the message to be transmitted. For simplicity, the carrier wave is modeled here as a simple sine wave written as

$$c(t) = C \cdot \cos(\omega_c t + \phi_c),$$

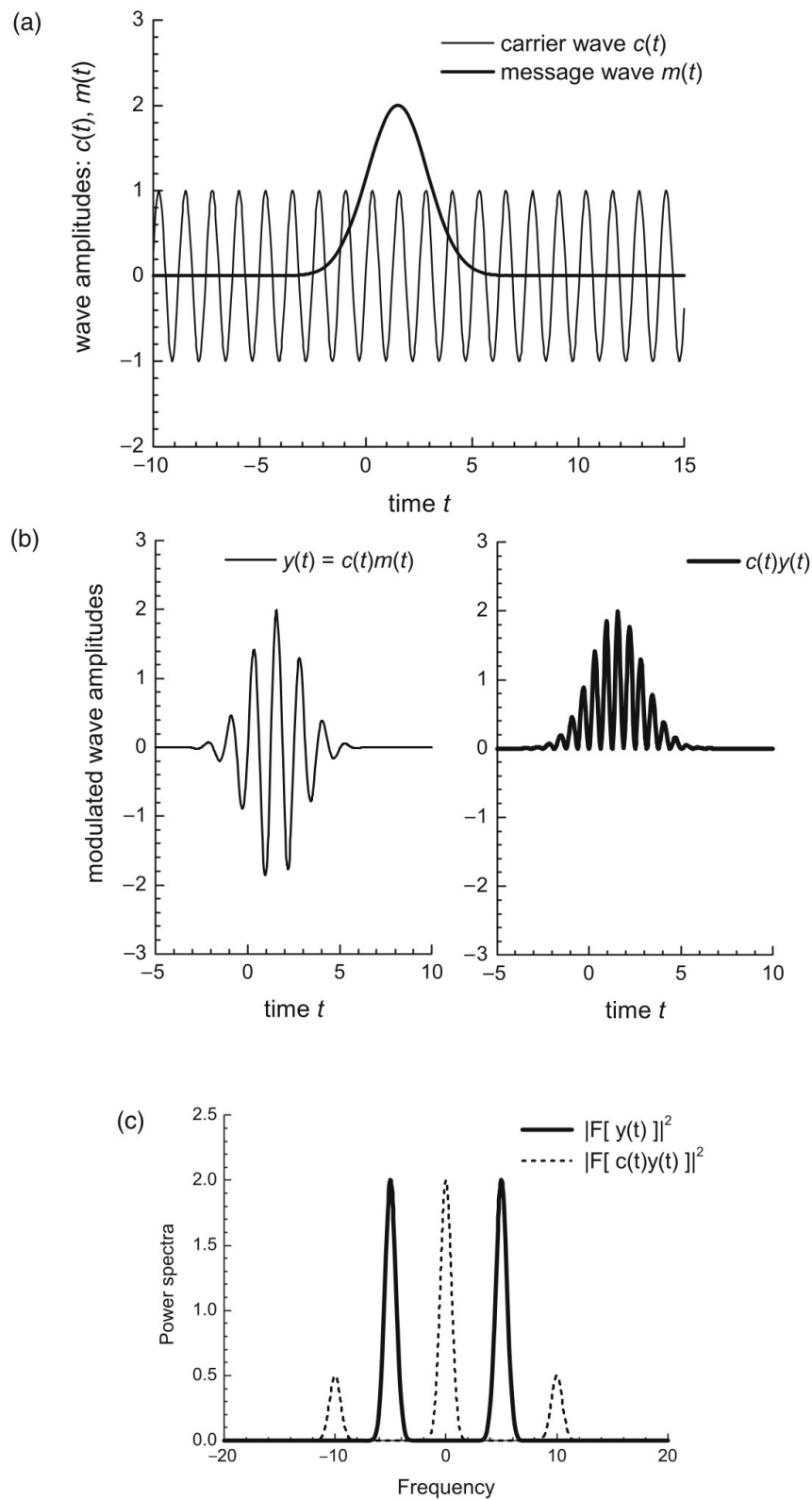
where the radio frequency (in Hertz) is given by  $\omega_c/(2\pi)$ .  $C$  and  $\phi_c$  are constants representing the carrier amplitude and the initial phase, respectively, and their values are set to 1 and 0. AM is then realized by determining the product:

$$y(t) = m(t) \cdot c(t),$$

whose Fourier transform  $Y(\omega)$  is expressed as

$$\begin{aligned} Y(\omega) &= \mathcal{F}[m(t)c(t)] \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} m(t) e^{-i\omega t} \frac{e^{i\omega_c t} + e^{-i\omega_c t}}{2} dt \\ &= \frac{1}{2} [M(\omega + \omega_c) + M(\omega - \omega_c)]. \end{aligned} \tag{12.49}$$

Here  $M(\omega)$  is the Fourier transform of  $m(t)$ .



**Fig. 12.2.** *Top:* A carrier wave  $c(t) = \sin(\omega_c t)$  with  $\omega_c = 5.0$  and a message wave  $m(t) = 2 \exp[-(t - t_0)^2/4]$  with  $t_0 = 1.5$ . *Middle:* The products  $c(t)m(t) \equiv y(t)$  and  $c^2(t)m(t)$ . *Bottom:* The power spectra  $|\mathcal{F}[y(t)]|^2$  and  $|\mathcal{F}[c(t)y(t)]|^2$

The result in (12.49) implies that the modulated signal  $y(t)$  has two groups of components: one at positive frequencies (centered at  $+\omega_c$ ) and one at negative frequencies (centered at  $-\omega_c$ ). Figure 12.2 illustrates a carrier wave  $c(t) = \sin(\omega_c t)$  with  $\omega_c = 5.0$ , a message wave  $m(t) = 2 \exp[-(t - t_0)^2/4]$  with  $t_0 = 1.5$ , and the power spectrum of  $y(t) = c(t)m(t)$  [i.e.,  $\omega$ -dependence of  $Y(\omega)$ ] described by (12.49), together with the associated message wave  $m(t)$ . The frequency shift from  $\omega$  to  $\omega \pm \omega_c$ , which is clearly evident, facilitates the tuning of the frequency of the transmitted signal to the desired value. We are concerned only with positive frequencies. The negative ones are mathematical artifacts that carry no additional information.

In order to reproduce the original signal  $m(t)$  from the modulated one  $y(t)$ , it is sufficient to multiply  $c(t)$  by  $y(t)$  and follow that with a filtering process. The Fourier transform of the product  $c(t)y(t)$  is given as

$$\begin{aligned}\mathcal{F}[c(t)y(t)] &= \mathcal{F}[m(t) \cos^2(\omega_c t)] \\ &= \frac{M(\omega)}{2} + \frac{1}{4} [M(\omega + 2\omega_c) + M(\omega - 2\omega_c)].\end{aligned}$$

We pick up the first term in the last expression and take its inverse transform, thus obtaining  $\mathcal{F}^{-1}[M(\omega)] = m(t)$ .

## Laplace Transformation

**Abstract** Using the Laplace transform for the mathematical description of a physical system considerably simplifies the analysis of its behavior. Many useful applications and formulas related to Laplace transforms can be found in other textbooks, but here we focus on the theoretical background, particularly, on the convergence properties of the various forms of Laplace transforms. It is important to note that a Laplace transform exists only if the corresponding improper integral, known as the Laplace integral, converges. Hence, the convergence of the improper integral must be confirmed prior to discussing the Laplace transform of a given function. Thus we devote a portion of this chapter to an analysis of the conditions necessary for the convergence of Laplace integrals, in contrast to the standard literature that deals primarily with the practical applications of Laplace transforms.

### 13.1 Basic Operations

#### 13.1.1 Definitions

The **Laplace transformation** associates a function  $f(x)$  of a real variable  $x$  with a suitable function  $F(s)$  of a complex variable  $s$ . This correspondence is essentially a reciprocal one-to-one and often allows us to replace a given complicated function by a simpler one. The advantage of this operation manifests particularly in applications to problems of **linear differential equations** (see Chap. 15). We shall see that the Laplace transformation allows us to reduce a linear differential equation of  $f(x)$  to a certain simple algebraic equation of  $F(s)$ , which yields solutions of the original differential equations more readily than other techniques. Furthermore, it turns out that this reduction method can be extended to systems of differential equations (ordinary and partial) as well as to integral equations, which enhances the importance of studying and understanding the Laplace transform.

To begin with, we define the Laplace transformation operator  $L$  that maps a function  $f(x)$  to a corresponding function  $F(s)$ :

♠ **Laplace transformation:**

The (one-sided) **Laplace transformation**, denoted by the operator  $L$ , is defined by

$$L[f(x)] = \int_0^{\infty} e^{-sx} f(x) dx = F(s), \quad (13.1)$$

which associates an image function  $F(s)$  of the complex variable  $s = \sigma + i\omega$  with a single-valued function  $f(x)$  ( $x$  real) such that the integral (13.1) exists.

♠ **Laplace integral:**

The integral given in (13.1) is called the **Laplace integral**. If the Laplace integral exists for a given  $f(x)$ , the image function  $F(s)$  is called the (one-sided) **Laplace transform** of  $f(x)$ .

It is important to keep in mind the difference between the Laplace *integral* and the Laplace *transform*. Namely, the Laplace transform exists only when the Laplace integral exists (i.e., converges). Convergence properties of Laplace integrals are determined by the value of  $s$  and the feature of the function  $f(x)$ , which is discussed fully in Sect. 13.3. In the meantime, we assume that  $f(x)$  is a function that allows the Laplace integral to converge for certain  $s$ .

### 13.1.2 Several Remarks

Below are several important remarks regarding the properties of the Laplace transform (13.1).

1. The definition (13.1) states that for a given  $F(s)$ , there is at most one *continuous* function  $f(x)$ . Nevertheless, it does not determine a unique  $f(x)$  because if  $f(x)$  in (13.1) were altered at a finite number of isolated points,  $F(s)$  would remain unchanged, as such discontinuous points make no contribution to the integral. For this reason, we assume in the remainder of this chapter that  $f(x)$  is continuous except at isolated points.
2. In order for the integral (13.1) to exist, any discontinuity of the integrand inside the interval  $(0, \infty)$  must be a finite jump so that there are right-hand and left-hand limits at those discontinuous points. An exception is a discontinuity at  $x = 0$  (if it exists); for instance, the function  $f(x) = 1/\sqrt{x}$  diverges at  $t = 0$  but the integral (13.1) exists.
3. The **inverse Laplace transform** of  $F(s)$  is a function  $f(x)$  such that  $L[f(x)] = F(s)$ . Hence, the operation of taking an inverse Laplace transform is denoted by  $L^{-1}$  and we have

$$L^{-1}[F(s)] = f(x).$$

This expression implies the possibility of dealing with the operators  $L$  and  $L^{-1}$  algebraically, just as the equation  $ax = y$  can be rewritten as  $x = a^{-1}y$ . At this point, it is not clear as to how the inverse operation  $L^{-1}$  is to be performed, but actual manipulations are discussed in detail in Sect. 13.4.2.

4. Not every function  $F(s)$  has an inverse Laplace transform. A sufficient condition for  $F(s)$  to have its inverse transform is presented in Sect. 13.4.2.

### 13.1.3 Significance of Analytic Continuation

Observe that the Laplace integral (13.1) involves a complex-valued term  $e^{-sx}$  in its integrand, which makes it difficult to employ the standard methods of integration that are applicable to real integrands. One way to proceed would be to use the equation  $e^{-sx} = e^{-\sigma x} \cos \omega x - ie^{-\sigma x} \sin \omega x$ , which yields two real integrands. This is, however, more complicated than necessary. An easier method is to make use of the following theorem, which is verified in Sect. 13.3.7:

♠ **Analytic property of Laplace transform:**

The Laplace transform  $F(s)$ , which is a complex-valued function of a complex variable  $s$ , is an **analytic function** in a region of  $\operatorname{Re}(s) > \sigma_c$  with a specific real number  $\sigma_c$ .

■ *Remark.* Just at  $\operatorname{Re}(s) = \sigma_c$ , however, no general conclusion can be drawn.

This theorem states that once the value of  $F(\sigma)$  on the real axis is known,  $F(s)$  on an arbitrary point of the complex plane can be obtained by simply replacing  $\sigma$  by  $s$ . This replacement is based on an **analytic continuation** from the semi-infinite line of the real axis,  $\sigma > \sigma_c$ , to the right half of the  $s$ -plane,  $\operatorname{Re}(s) > \sigma_c$ , which is why we can perform the integration (13.1) as if  $s$  were a real variable. Several examples given later clearly show the efficacy of identifying  $s$  as a real parameter.

At first glance, the formality of replacing  $\sigma$  by  $s$  amounts simply to a change in symbol. But, without analytic continuation, we could no longer regard our replacement from  $\sigma$  to  $s$  as a mere formality; i.e., the concept of analytic continuation lurks in the background.

■ *Remark.* In particular, those cases in which  $F(s)$  becomes multivalued cannot be treated without paying heed in detail to the difference between  $\sigma$  and  $s$ . The latter issue regarding multivalued  $F(s)$  is discussed in Sect. 13.2.5.

### 13.1.4 Convergence of Laplace Integrals

Emphasis is placed on the fact that the Laplace integral (13.1) may or may not exist depending on the value of  $s$  as well as the nature of  $f(x)$ . A sufficient condition for the Laplace integral to converge is that the real component of  $s$ ,  $\operatorname{Re}(s)$ , is greater than a specific value. This intuitively follows from the definition (13.1) that says if the integral (13.1) exists for

$$s_0 = \sigma_0 + i\omega_0,$$

then the integral also exists for every  $s$  such that  $\operatorname{Re}(s) > \sigma_0$ , since in the latter case

$$|e^{-sx}| < |e^{-s_0x}| = e^{-\sigma_0x}.$$

This is stated rigorously in the theorem below.

♠ **Convergence of Laplace integrals:**

If the Laplace integral

$$\int_0^\infty f(x)e^{-sx}dx \tag{13.2}$$

converges for  $\operatorname{Re}(s) = \sigma_0$ , then it also converges for  $\operatorname{Re}(s) > \sigma_0$ .

The proof is given in Sect. 13.3.4. This theorem implies the existence of a specific real number  $\sigma_c$  such that the integral (13.2) converges for  $\operatorname{Re}(s) > \sigma_c$  and diverges for  $\operatorname{Re}(s) < \sigma_c$  (see Fig. 13.1). The number  $\sigma_c$  is called the **abscissa of convergence** of the Laplace integral, whose value depends on the nature of the function  $f(x)$ . With this notation, we say that **the region of convergence of the Laplace integral** is a half-plane to the right of  $\operatorname{Re}(s) = \sigma_c$ . This region of convergence is of course identified with the defining region of the Laplace transform  $F(s)$ .

**Remark.** By definition,  $\sigma_c$  may take  $-\infty$  (or  $\infty$ ), which means that the integral (13.2) converges (or diverges) for all  $\sigma$ .

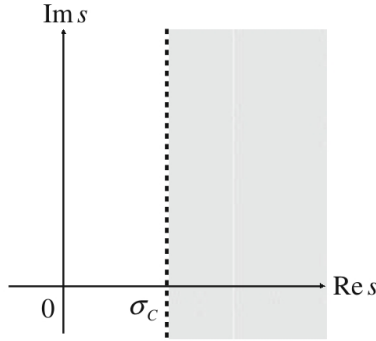
*Examples* Set  $f(x) = 1$  for every  $x \geq 0$ . Then

$$\begin{aligned} L[f(x)] &= \int_0^\infty e^{-sx}dx = \lim_{X \rightarrow \infty} \int_0^X s^{-sx}dx \\ &= \lim_{X \rightarrow \infty} \left[ \frac{s^{-sx}}{-s} \right]_0^X = \frac{1}{s} - \frac{1}{s} \lim_{X \rightarrow \infty} e^{-sX}. \end{aligned}$$

Hence, we have

$$L[f(x)] = \frac{1}{s} \text{ for } s > 0.$$

For  $s \leq 0$ , the integral does not converge. This indicates that in this case  $\sigma_c = 0$ .



**Fig. 13.1.** The abscissa of convergence  $\sigma_c$  to the right of which the Laplace integral converges

### 13.1.5 Abscissa of Absolute Convergence

When the Laplace integral converges in the ordinary sense, it might converge *absolutely* in part or in all of its converging region. (Remember that the conditions for absolute convergence are more stringent than those for ordinary convergence). This leads us to define an abscissa of *absolute* convergence as follows:

#### ♠ Abscissa of absolute convergence:

Suppose that the Laplace integral (13.2) converges absolutely for  $\text{Re}(s) = \sigma_0$  as

$$\int_0^\infty |f(x)e^{-sx}| dx = \int_0^\infty |f(x)|e^{-\sigma_0 x} dx < \infty. \quad (13.3)$$

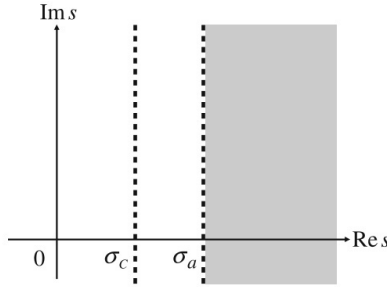
The greatest lower bound  $\sigma_a$  of such a  $\sigma_0$  that satisfies (13.3) is called the **abscissa of absolute convergence** of the Laplace integral (13.2).

Thus once  $\sigma_a$  is determined, we say that the integral (13.2) converges absolutely for  $\sigma > \sigma_a$ , does not converge absolutely for  $\sigma < \sigma_a$ , and may or

may not converge absolutely at  $\sigma = \sigma_a$ . Since absolute convergence implies ordinary convergence, it is clear that

$$\sigma_c \leq \sigma_a.$$

The following example shows that  $\sigma_a$  does not generally coincide with  $\sigma_c$  (see Fig. 13.2).



**Fig. 13.2.** The abscissa of convergence  $\sigma_c$  and the abscissa of *absolute* convergence  $\sigma_a$

*Example*  $f(x) = e^x \sin e^x$

Set  $u = e^x$ ; then we have

$$F(s) = \int_0^\infty e^{-sx} e^x \sin e^x dx = \int_1^\infty \frac{\sin u}{u^s} du.$$

The integral converges absolutely for  $\text{Re}(s) = \sigma > 1$ , converges conditionally for  $0 < \sigma \leq 1$ , and diverges for  $\sigma = 0$ . Hence, we have

$$\sigma_c = 0 \quad \text{and} \quad \sigma_a = 1,$$

which clearly indicates that in this case  $\sigma_c \neq \sigma_a$ .

### 13.1.6 Laplace Transforms of Elementary Functions

Let us evaluate the Laplace transforms  $F(s)$  of several classes of elementary functions. We treat the complex variable  $s$  as if it were real, bearing in mind that this formalism is based on the **analyticity** of  $F(s)$ , as discussed in Sect. 13.1.3. The defining region of each  $F(s)$  is found on the right-hand side of the equation in question.

1.  $f(x) = x^n$ , where  $n$  is a positive integer.

Integrating by parts, we have

$$\begin{aligned}
F(s) = L[x^n] &= \int_0^\infty x^n e^{-sx} dx \\
&= \left[ \frac{-x^n e^{-sx}}{s} \right]_0^\infty + \frac{n}{s} \int_0^\infty t^{n-1} e^{-sx} dx.
\end{aligned} \tag{13.4}$$

Since  $s > 0$  and  $n > 0$ , the first term in the last expression of (13.4) vanishes. Iteration of this process yields

$$L[x^n] = \frac{n(n-1)(n-2)\cdots 2 \cdot 1}{s^n} L[t^0] = \frac{n!}{s^{n+1}}$$

since  $L[t^0] = L[1] = 1/s$ . As a result, we have

$$F(s) = L[x^n] = \frac{n!}{s^{n+1}} \quad (\sigma > 0).$$

**2.**  $f(x) = e^{ax}$ , where  $a$  is a real constant.

$$F(s) = L[e^{ax}] = \int_0^\infty e^{-sx} e^{ax} dx = \frac{1}{s-a} \quad (\sigma > a).$$

**3.**  $f(x) = \sin ax$ , where  $a$  is a real constant.

Integrating by parts twice, we obtain

$$\begin{aligned}
F(s) = L[\sin ax] &= \int_0^\infty e^{-sx} \sin ax dx \\
&= \left[ -\frac{e^{-sx}}{a} \cos ax \right]_0^\infty + \frac{1}{a} \int_0^\infty (-s) e^{-sx} \cos ax dx \\
&= \frac{1}{a} - \frac{s}{a} \left\{ \left[ \frac{e^{-sx}}{a} \sin ax \right]_0^\infty + \frac{s}{a} \int_0^\infty e^{-sx} \sin ax dx \right\} \\
&= \frac{1}{a} - \frac{s^2}{a^2} F(s),
\end{aligned}$$

where we have used the fact that as  $s$  is positive,  $e^{-sx} \rightarrow 0$  as  $x \rightarrow \infty$ , whereas  $\sin ax$  and  $\cos ax$  are bounded as  $x \rightarrow \infty$ . Eventually, we set to

$$F(s) = L[\sin ax] = \frac{a}{s^2 + a^2}. \quad (\sigma > 0).$$

In a similar manner, we obtain

$$L[\cos ax] = \int_0^\infty e^{-sx} \cos ax dx = \frac{s}{s^2 + a^2} \quad (\sigma > 0).$$

4.  $f(x) = \cosh ax$ , where  $a$  is a real constant.

Using the linearity property of the Laplace transform operator  $L$ , we obtain

$$\begin{aligned} L[\cosh ax] &= L\left[\frac{e^{ax} + e^{-ax}}{2}\right] = \frac{1}{2}L[e^{ax}] + \frac{1}{2}L[e^{-ax}] \\ &= \frac{1}{2}\left(\frac{1}{s-a} + \frac{1}{s+a}\right) = \frac{s}{s^2 - a^2} \quad (\sigma > |a|). \end{aligned}$$

## Exercises

1. Show the linearity of the Laplace transformation operator  $L$ .

**Solution:** It follows from the definition of the operator  $L$  that

$$\begin{aligned} L[c_1 f(x) + c_2 g(x)] &= \int_0^\infty e^{-sx} \{c_1 f(x) + c_2 g(x)\} dx \\ &= c_1 \int_0^\infty e^{-sx} f(x) dx + c_2 \int_0^\infty e^{-sx} g(x) dx \\ &= c_1 L[f(x)] + c_2 L[g(x)], \end{aligned}$$

where  $c_1$  and  $c_2$  are arbitrary constants. This clearly shows the linearity of the operator  $L$ . ♣

2. Find the Laplace transform of the function,

$$f(x) = \begin{cases} 0, & 0 \leq x < c, \\ 1, & x \geq c. \end{cases}$$

**Solution:**  $L[f(x)] = \int_0^\infty e^{-sx} f(x) dx = \int_c^\infty e^{-sx} dx = e^{-cs}/s \quad (\sigma > 0).$

♣

3. Show that if  $f(x)$  is real and  $F(x) = L[f(x)]$  is single-valued, then  $F(s)$  is real.

**Solution:** Set  $s = \sigma > \sigma_c$  in the equation  $F(s) = \int_0^\infty f(x)e^{-sx} dx$ .

Then the integrand  $f(x)e^{-\sigma x}$  is real, so  $F(\sigma)$  is real. This establishes that  $F(s)$  is real on the real axis to the right of the point  $s = \sigma_c$ . In view of analytic continuation, therefore,  $F(s)$  is a real-valued analytic function. ♣

## 13.2 Properties of Laplace Transforms

### 13.2.1 First Shifting Theorem

In physical applications, we are sometimes required to calculate the Laplace transform of functions multiplied by exponential factors such as

$$e^{-ax} f(x),$$

where  $a$  is real or complex. This kind of problem can be simplified by applying the theorem below.

♠ **The first shifting theorem:**

If  $F(s) = L[f(x)]$  for  $\sigma > \sigma_c$ , then

$$F(s + a) = L[e^{-ax} f(x)]$$

for  $\sigma > \sigma_c - \operatorname{Re}(a)$ , where  $a$  is real or complex.

*Proof* Suppose  $\sigma_c$  to be the abscissa of convergence for  $F(s)$ . Then the integral

$$\int_0^\infty e^{-ax} f(x) e^{-sx} dx = \int_0^\infty f(x) e^{-(s+a)x} dx \quad (13.5)$$

clearly converges for  $\operatorname{Re}(s + a) > \sigma_c$ . Observe that the integral on the right-hand side of (13.5) is an expression for  $F(s + a)$ . Thus we have the general result:

$$L[e^{-ax} f(x)] = F(s + a),$$

where  $F(s) = L[f(x)]$ . ♣

The above theorem states that if we know the Laplace transform of any function, the transform of that function multiplied by an exponential can immediately be obtained by a simple *shift* (or translation) in the  $s$  variable.

*Examples* 1. The first shifting theorem tells that

$$L[e^{-ax} x^n] = \frac{n}{(s + a)^{n+1}}, \quad \sigma > -a,$$

since

$$L[x^n] = \frac{n}{s^{n+1}}, \quad \sigma > 0.$$

2. Similarly, it follows from the first shifting theorem that

$$L[e^{-ax} \sin bt] = \frac{b}{(s+a)^2 + b^2}, \quad \sigma > -a,$$

where we use the fact that

$$L[\sin at] = \frac{a}{\sigma^2 + a^2}.$$

### 13.2.2 Second Shifting Theorem

For the next case, assume again that a function  $f(x)$  has a transform  $F(s)$  and consider a shift in the  $x$  variable from  $x$  to  $x - x_0$ , where  $x_0$  is a positive constant. Stipulating that the new function be zero for  $x < x_0$ , it can be written

$$f(x - x_0)\theta(x - x_0), \quad (13.6)$$

where

$$\theta(x) = \begin{cases} 0, & x < 0, \\ 1, & x > 0. \end{cases}$$

The Laplace transform of the shifted function (13.6) is thus represented by the integral

$$\int_0^\infty f(x - x_0)\theta(x - x_0)e^{-sx} dx = \int_{x_0}^\infty f(x - x_0)e^{-sx} dx.$$

Now we change the variable of integration to  $t' = x - x_0$ , which gives us

$$L[f(x - x_0)\theta(x - x_0)] = e^{-sx_0} \int_0^\infty f(t')e^{-st'} dx' = e^{-sx_0} F(s).$$

The result is stated formally below.

♠ **The second shifting theorem:**

If  $F(s) = L[f(x)]$  for  $\sigma > \sigma_c$ , then

$$e^{-sx_0} F(s) = L[f(x - x_0)\theta(x - x_0)]$$

for  $\sigma > \sigma_c$ , where  $\theta(x)$  is a unit step function and  $T$  is a real and positive constant.

*Examples* Consider the Laplace transform of the function

$$f(x) = \begin{cases} 0 & (x < 0), \\ 1/a & (0 \leq x < a), \\ 0 & (x \geq a). \end{cases}$$

Using the step function, we express it as

$$f(x) = \frac{\theta(x) - \theta(x-a)}{a}.$$

Hence, it follows from the second shifting theorem that

$$L[f(x)] = \frac{L[1] - e^{-as}L[1]}{a} = \frac{1 - e^{-as}}{as}.$$

Note that in view of **l'Hôpital's rule**, when  $a \rightarrow 0$ ,  $L[f(x)] = 1$ . The latter result means that the Laplace transform of  $f(x)$  equals 1.

### 13.2.3 Laplace Transform of Periodic Functions

We now consider the Laplace transform of a periodic function  $f(x)$  of period  $\lambda$ , i.e.,  $f(x + \lambda) = f(x)$ . Assuming that the  $f(x)$  is piecewise continuous, we have by definition

$$\begin{aligned} L[f(x)] &= \int_0^{\infty} e^{-sx} f(x) dx \\ &= \int_0^{\lambda} e^{-sx} f(x) dx + \int_{\lambda}^{2\lambda} e^{-sx} f(x) dx + \int_{2\lambda}^{3\lambda} e^{-sx} f(x) dx + \cdots \end{aligned}$$

On the right-hand side, let  $x = u + \lambda$  in the second integral,  $x = u + 2\lambda$  in the third integral, and so on. We then set

$$\begin{aligned} L[f(x)] &= \int_0^{\lambda} e^{-sx} f(x) dx + \int_0^{\lambda} e^{-s(u+\lambda)} f(u + \lambda) du \\ &\quad + \int_0^{\lambda} e^{-s(u+2\lambda)} f(u + 2\lambda) du + \cdots \end{aligned}$$

From hypothesis,  $f(u + \lambda) = f(u)$ ,  $f(u + 2\lambda) = f(u)$ , etc. Replacing the dummy variable  $u$  by  $x$  yields

$$\begin{aligned} L[f(x)] &= (1 + e^{-s\lambda} + e^{-2s\lambda} + \cdots) \int_0^{\lambda} e^{-sx} f(x) dx \\ &= \frac{1}{1 - e^{-s\lambda}} \int_0^{\lambda} e^{-sx} f(x) dx. \end{aligned} \tag{13.7}$$

Once we introduce the function

$$f_0(x) = \begin{cases} f(x), & 0 \leq x \leq \lambda, \\ 0, & \text{otherwise,} \end{cases}$$

equation (13.7) becomes

$$L[f(x)] = \frac{F_0(s)}{1 - e^{-s\lambda}},$$

where

$$F_0(s) = \int_0^\infty e^{-sx} f_0(x) dx = \int_0^\lambda e^{-sx} f(x) dx,$$

So we have proven the following result:

♠ **Laplace transform of a periodic function:**

If  $f(x)$  is a periodic function of period  $\lambda$ , its Laplace transform is given by

$$F(s) = \frac{F_0(s)}{1 - e^{-s\lambda}}, \quad (13.8)$$

where

$$F_0(s) = \int_0^\lambda e^{-sx} f(x) dx.$$

*Examples* Consider the Laplace transform of the periodic square wave described by  $f(x + 2\lambda) = f(x)$  with

$$f(x) = \begin{cases} 1 & (0 < x < \lambda), \\ -1 & (\lambda < x < 2\lambda). \end{cases}$$

From (13.8), we obtain

$$\begin{aligned} F(s) &= \frac{1}{1 - e^{-2s\lambda}} \int_0^{2\lambda} e^{-sx} f(x) dx \\ &= \frac{1}{1 - e^{-2s\lambda}} \left( \int_0^\lambda - \int_\lambda^{2\lambda} \right) e^{-sx} f(x) dx \\ &= \frac{(1 - e^{-s\lambda})^2}{s(1 - e^{-2s\lambda})} = \frac{1}{s} \tanh\left(\frac{s\lambda}{2}\right). \end{aligned}$$

### 13.2.4 Laplace Transform of Derivatives and Integrals

The Laplace transform of derivatives is a most important issue in terms of applications for solving differential equations. We shall see below that through the transform, certain kinds of differential equations are reduced to algebraic equations that are easy to manipulate.

♠ **Laplace transform of derivatives:**

If  $F(s) = L[f(x)]$  for  $\sigma > \sigma_c$  and if

$$\lim_{t \rightarrow \infty} e^{-sx} f(x) = 0 \quad \text{for } \sigma > \sigma_c, \quad (13.9)$$

then we have

$$L[f'(x)] = sF(s) - f(0). \quad (13.10)$$

*Proof* Integration by parts yields

$$\begin{aligned} L[f'(x)] &= \int_0^\infty e^{-sx} f'(x) dx \\ &= [e^{-sx} f(x)]_0^\infty - \int_0^\infty (-s)e^{-sx} f(x) dx. \end{aligned} \quad (13.11)$$

The second term on the right-hand side of (13.11) converges to  $sF(s)$  for  $\sigma > \sigma_c$ . In addition, the first term reads  $f(0)$  from the hypothesis of (13.9). Thus for  $\sigma > \sigma_c$ , we obtain (13.10). ♣

This result can be extended to cases of higher derivatives.

♠ **Laplace transform of higher derivatives:**

Suppose  $f(x)$  to be such that  $f^{(n-1)}(x)$  is continuous. If  $F(s) = L[f(x)]$  for  $\sigma > \sigma_c$  and if

$$\lim_{t \rightarrow \infty} e^{-sx} f^{(k)}(x) = 0$$

for  $k = 0, 1, \dots, n-1$  and  $\sigma > \sigma_c$ , then

$$L[f^{(n)}(x)] = s^n F(s) - \sum_{k=1}^n s^{n-k} f^{(k-1)}(0).$$

The above theorem is central to the use of the Laplace transform for solving differential equations with specified initial conditions (i.e., **initial value problems**).

♠ **Laplace transform of integrals:**

If  $g(x) = \int_0^x f(u)du$ ,  $L[f(x)] = F(s)$  and if

$$\lim_{t \rightarrow \infty} e^{-sx} g(x) = 0,$$

then

$$L[g(x)] = \frac{F(s)}{s}.$$

*Proof* From hypothesis, we have  $g(0) = 0$  and  $g'(x) = f(x)$ , and thus

$$L[g'(x)] = L[f(x)].$$

The left-hand side becomes

$$L[g'(x)] = sL[g(x)] - g(0) = sL[g(x)].$$

As a result, we obtain

$$L[g(x)] = \frac{1}{s} L[f(x)] = \frac{F(s)}{s}. \quad \clubsuit$$

### 13.2.5 Laplace Transforms Leading to Multivalued Functions

Some care should be taken when the Laplace transform results in a **multi-valued function**. A typical example is the transform of the function

$$f(x) = \frac{1}{\sqrt{x}} \quad (x \geq 0). \quad (13.12)$$

Although this function has a singularity at  $x = 0$ , the **improper integral** having a real integrand,

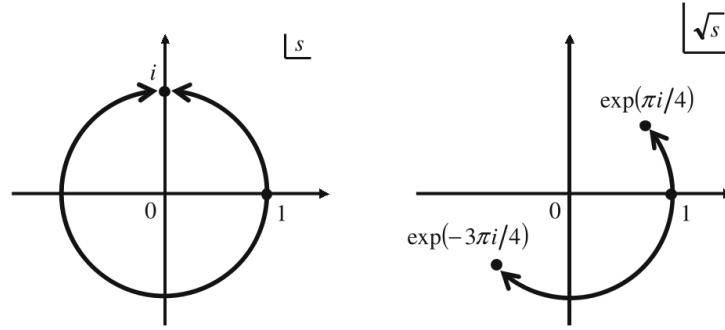
$$\int_0^\infty \frac{e^{-\sigma x}}{\sqrt{x}} dx, \quad (13.13)$$

converges for  $\sigma > 0$ . In what follows, we first evaluate the integral (13.13) and then **continue analytically** with the result to arrive at a suitable region of the complex  $s$ -plane where we can get a precise form of  $F(s)$ .

The integral (13.13) can be readily evaluated by setting  $\sigma x = u^2$ ; then it reads

$$\frac{2}{\sqrt{\sigma}} \int_0^\infty e^{-u^2} du = \frac{\sqrt{\pi}}{\sqrt{\sigma}}. \quad (13.14)$$

Now we would like to continue analytically to take the result (13.14) to the complex  $s$ -plane. At first glance, it suffices to replace  $\sqrt{\sigma}$  by  $\sqrt{s}$  symbolically. However, this is not sufficient because the function  $\sqrt{s}$  is double-valued (e.g., when  $s = i = e^{\pi i/2}$ ,  $\sqrt{s}$  may take the two distinct values:  $e^{\pi i/4}$  and  $e^{-3\pi i/4}$ ;



**Fig. 13.3.** The double-valuedness of the function  $\sqrt{s}$

(see Fig. 13.3). Thus we have two possible choices (i.e., two sheets of **Riemann surfaces**) when performing analytic continuation from the real single-valued function  $\sqrt{\sigma}$  to the complex double-valued function  $\sqrt{s}$ . We go into only one sheet of Riemann surface, the choice being the one on which the points of  $\sqrt{\sigma}$  are situated [i.e., the right half of the whole  $s$ -plane, expressed by  $\operatorname{Re}(s) > 0$ ]. With this convention, we arrive at the result

$$F(s) = L \left[ \frac{1}{\sqrt{x}} \right] = \frac{\sqrt{\pi}}{\sqrt{s}}, \quad (13.15)$$

where the symbol  $\sqrt{s}$  implies the single-valued branch mentioned above.

**Remark.** If the above case had been treated throughout with the variable  $s$  retained, the formal variable change would have led to the factor  $1/\sqrt{s}$  as in (13.15). However, we would not then have a clear meaning for  $\sqrt{s}$ ; i.e., there would be no way to determine which branch is to be taken.

## Exercises

1. Show that

$$\lim_{x \rightarrow +0} f(x) = \lim_{s \rightarrow \infty} sF(s)$$

and

$$\lim_{x \rightarrow \infty} f(x) = \lim_{s \rightarrow 0} sF(s),$$

where the Laplace integral  $L[f(x)] = F(s)$  converges for  $\sigma \geq 0$ .

**Solution:** Take the limits  $s \rightarrow \infty$  on both sides of equation

$$\int_0^\infty f'(x) e^{-sx} dx = sF(s) - f(0). \quad (13.16)$$

Then we have  $0 = \lim_{s \rightarrow \infty} sF(s) - f(0+)$ , which gives us our first result. Moreover, in the limit  $s \rightarrow 0$ , the left-hand side of (13.16) reads  $\int_0^\infty f'(x) dx = \lim_{x \rightarrow \infty} f(x) - f(0+)$ , so that we set our second result. ♣

2. Find the transform of the function

$$f(x) = \sqrt{t^k},$$

where  $k \geq 1$  and is an odd integer.

**Solution:** This function gives convergence for  $\sigma > 0$ . Integration by parts yields a general recurrence equation:

$$\int_0^\infty \sqrt{t^k} e^{-\sigma x} dx = - \left[ \frac{\sqrt{t^k} e^{-\sigma x}}{\sigma} \right]_0^\infty + \frac{k}{2\sigma} \int_0^\infty \sqrt{t^{k-2}} e^{-\sigma x} dx.$$

Since  $k \geq 1$ , the lower limit can be used in the first term on the right-hand side (and thus the integral exists). The result can be stated as

$$L[\sqrt{t^k}] = \frac{k}{2s} L[\sqrt{t^{k-2}}] \quad \text{where } k \geq 1 \text{ and odd.}$$

This yields a sequence of equations, starting with  $\sqrt{t^{-1}}$ , that is obtained from (13.15). Consequently, we have

$$L\left[\frac{1}{\sqrt{x}}\right] = \frac{\sqrt{\pi}}{\sqrt{s}}, \quad L[\sqrt{x}] = \frac{\sqrt{\pi}}{2\sqrt{s^3}}, \quad L[\sqrt{x^3}] = \frac{3\sqrt{\pi}}{4\sqrt{s^5}}, \quad \dots$$

$$L[\sqrt{x^k}] = \frac{(k+1)! \sqrt{\pi}}{2^{k+1}[(k+1)/2]! \sqrt{s^{k+2}}}.$$

In these general equations, the root of a power of  $s$  is always interpreted as being on the sheet of the Riemann surface on which the values of  $\sqrt{\sigma^{k+2}}$  are found. ♣

## 13.3 Convergence Theorems for Laplace Integrals

### 13.3.1 Functions of Exponential Order

The Laplace integral is **improper** by virtue of an infinite limit of integration, as shown clearly by

$$\int_0^\infty f(x) e^{-sx} dx = \lim_{R \rightarrow \infty} \int_0^R f(x) e^{-sx} dx. \quad (13.17)$$

This improper integral can be identified with the Laplace transform  $F(s)$  *only* when it converges for the values of  $s$  in question. Therefore, it is important to clarify the conditions under which the Laplace integral converges. As a first step in addressing this issue, we introduce a new class of functions:

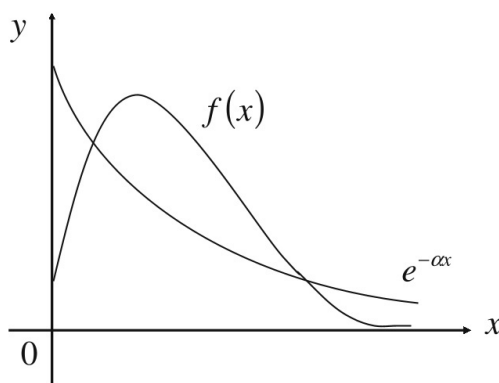
♠ **Functions of exponential order:**

A function  $f(x)$  is said to be of **exponential order**  $\alpha_0$  if there is a real number  $\alpha_0$  such that

$$\lim_{x \rightarrow \infty} f(x)e^{-\alpha x} = 0 \quad \text{for any } \alpha > \alpha_0, \quad (13.18)$$

and with the limit not existing when  $\alpha < \alpha_0$ .

See Fig. 13.4 for the decaying behavior of a function  $f(x)$  of exponential order  $\alpha$ . Note that condition (13.18) is not necessarily satisfied at  $\alpha = \alpha_0$ . The order number  $\alpha_0$  may take  $-\infty$  if  $f(x)$  is identically zero beyond some finite value of  $x$ .



**Fig. 13.4.** Decaying behavior of a function  $f(x)$  of exponential order  $\alpha$

*Examples* **1.** The function  $f(x) = x^3$  is of exponential order zero. To see this, it suffices to check whether or not

$$\lim_{x \rightarrow \infty} (e^{-\alpha x} x^3) \quad (13.19)$$

exists. If  $\alpha > 0$ , then **l'Hôpital's rule** gives

$$\lim_{x \rightarrow \infty} \frac{x^3}{e^{\alpha x}} = \lim_{x \rightarrow \infty} \frac{6}{\alpha^3 e^{\alpha x}} = 0.$$

In contrast, when  $\alpha < 0$ , (13.19) obviously diverges. Therefore  $x^3$  is of exponential order zero. In a similar manner, it can be shown that  $x^n$  for any integer  $n \geq 0$  is of exponential order zero.

**2.** The function  $f(x) = e^{cx}$  with any real constant  $c$  is of exponential order  $c$ , owing to the fact that

$$\lim_{x \rightarrow \infty} e^{cx} e^{-\alpha x} = 0$$

if and only if  $\alpha > c$ .

### 13.3.2 Convergence for Exponential-Order Cases

Suppose  $f(x)$  to be of exponential order  $\alpha_0$ . Then, we can show that the Laplace integral

$$\int_0^\infty f(x) e^{-sx} dx \quad (13.20)$$

converges absolutely whenever the real component of  $s$  is located within the range

$$\operatorname{Re}(s) = \sigma > \alpha_0. \quad (13.21)$$

Since absolute convergence implies ordinary convergence, the inequality (13.21) serves as a sufficient condition for the Laplace integral (13.20) to converge. This result is formally stated by the theorem below.

♠ **Theorem:** (= A sufficient condition for convergence for exponential-order cases)

If  $f(x)$  is of exponential order  $\alpha_0$ , then the Laplace integral  $\int_0^\infty f(x) e^{-sx} dx$  converges for

$$\operatorname{Re}(s) > \alpha_0.$$

(See also Fig. 13.5.)

*Proof* For any  $\sigma$  in the range of (13.21), we can pick a number  $\alpha_1$  such that

$$\alpha_0 < \alpha_1 < \sigma.$$

Since  $f(x)$  is of exponential order  $\alpha_0$ , we have

$$\lim_{x \rightarrow \infty} f(x) e^{-\alpha_1 x} = 0.$$

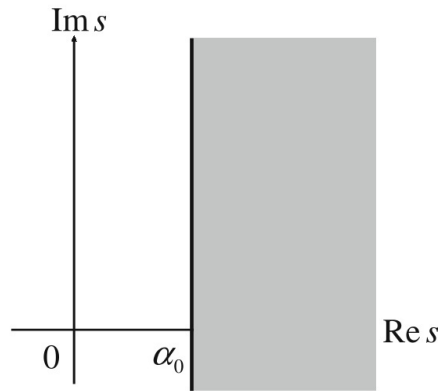
This implies that for any given small  $\varepsilon > 0$ , we can find an appropriate  $X$  such that

$$|f(x)| e^{-\alpha_1 x} < \varepsilon \quad \text{for any } x > X.$$

Hence, given any small  $\varepsilon > 0$ , there exists an  $X$  such that for  $A, A' > X$ ,

$$\begin{aligned} \int_A^{A'} |f(x)| e^{-\sigma x} dx &= \int_A^{A'} |f(x)| e^{-\alpha_1 x} e^{-(\sigma - \alpha_1)x} dx \\ &< \varepsilon \int_A^{A'} e^{-(\sigma - \alpha_1)x} dx, \end{aligned} \quad (13.22)$$

where the last integral in (13.22) converges to a finite value because  $\sigma > \alpha_1$ . This means that the leftmost integral in (13.22) can be made to approach zero by taking  $X$  sufficiently large. Thus in view of the **Cauchy's test for improper integrals** given in Sect. 3.4.4, the inequality (13.22) establishes the absolute (and thus ordinary) convergence of the integral (13.20) in the region  $\operatorname{Re}(s) > \alpha_0$ . ♣



**Fig. 13.5.** Converging region of the Laplace integral of a function of exponential order  $\alpha_0$

**Remark.** The above theorem provides a sufficient condition for the ordinary convergence of the Laplace integral. Hence, a given Laplace integral of the function of exponential order  $\alpha_0$  *must* converge for  $\operatorname{Re}(s) > \alpha_0$ , whereas it *may or may not* converge at  $\operatorname{Re}(s) \leq \alpha_0$ . For example,  $f(x) = \cos e^x$  gives  $\alpha_0 = 0$ , but the corresponding Laplace integral converges for  $\operatorname{Re}(s) > -1$ .

### 13.3.3 Uniform Convergence for Exponential-Order Cases

Next we examine the condition for *uniform* convergence. Here, uniform convergence means that the improper integral (13.20) as a function of  $s$  converges uniformly to  $F(s)$  over the whole defining region of the  $s$ -plane. To proceed, let  $\alpha_2$  be a number greater than  $\alpha_0$  and let  $\sigma$  be in the range

$$\alpha_0 < \alpha_2 \leq \sigma. \quad (13.23)$$

For any choice of  $\alpha_2$ , we can find a number  $\alpha_1$  such that

$$\alpha_0 < \alpha_1 < \alpha_2.$$

The relation (13.22) is again valid by use of  $\alpha_2$  instead of  $\alpha_1$ , as expressed by

$$\int_A^{A'} |f(x)| e^{-\sigma x} dx < \varepsilon \int_A^{A'} e^{-(\sigma - \alpha_2)x} dx.$$

Furthermore, by introducing  $\alpha_1$ , we can extend this inequality to

$$\int_A^{A'} |f(x)| e^{-\sigma x} dx < \varepsilon \int_A^{A'} e^{-(\alpha_1 - \alpha_2)x} dx.$$

Note that the last integral converges and is independent of  $\sigma$ . Therefore, in view of the **Weierstrass test for improper integrals** (see Sect. 3.4.4), the Laplace integral  $\int_0^\infty f(x) e^{-sx} dx$  converges uniformly for  $\operatorname{Re}(s) \geq \alpha_2 > \alpha_0$ . We have thus proved the following theorem:

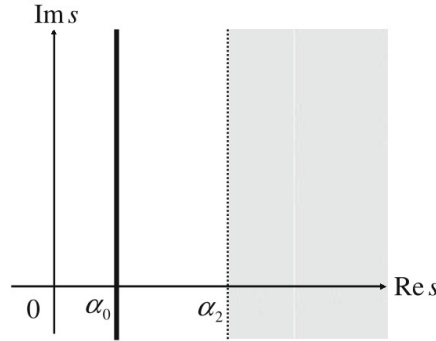
♠ **Theorem:** (= A sufficient condition for uniform convergence for exponential-order cases)

If  $f(x)$  is of exponential order  $\alpha_0$ , then the Laplace integral  $\int_0^\infty f(x) e^{-sx} dx$  converges uniformly to  $F(s) = L[f(x)]$  for

$$\operatorname{Re}(s) \geq \alpha_2 > \alpha_0.$$

(See also Fig. 13.6.)

Here, the constant  $\alpha_2$  emphasizes that the converging region guaranteed by this theorem is closed at the lower end.



**Fig. 13.6.** The region of uniform convergence associated with a function of exponential order  $\alpha_0$

**Remark.** It is important to remember that the above theorem gives only a *sufficient* condition for convergence of the Laplace integral. In fact, it is possible that some functions of exponential order allow their Laplace integrals to converge uniformly to the *left* of  $\alpha_0$ .

### 13.3.4 Convergence for General Cases

The previous two theorems tell a great deal about convergence of the Laplace integral for practical functions. On the other hand, for functions that are not of exponential order (but continuous within the integration interval), the following slightly different theorem applies.

♠ **Theorem:** (= A sufficient condition for convergence for general cases)

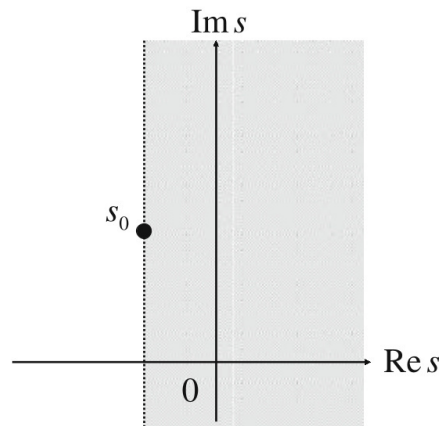
If the improper integral

$$\int_0^{\infty} f(x)e^{-sx} dx$$

converges for  $s = s_0$ , then it converges for  $\operatorname{Re}(s) > \operatorname{Re}(s_0)$  (see also Fig. 13.7).

*Proof* The proof requires an auxiliary function

$$g(x) = \int_x^{\infty} f(\tau)e^{-s_0\tau} d\tau, \quad (13.24)$$



**Fig. 13.7.** Converging region of the Laplace integral that converges for  $s = s_0$

where  $f(x)$  is assumed to satisfy the conditions given above. Since  $f(x)$  is continuous,  $g(x)$  is also continuous and thus its derivative is given by

$$g'(x) = -f(x)e^{-s_0x}.$$

In terms of  $g(x)$ , the Laplace integral can be written as

$$\int_0^\infty f(x)e^{-sx}dx = \int_0^\infty f(x)e^{-s_0x}e^{-wx}dx = -\int_0^\infty g'(x)e^{-wx}dx, \quad (13.25)$$

where we have set  $w = s - s_0$ .

We now examine sufficient conditions for the rightmost integral in (13.25) to converge. **Cauchy's test for improper integrals** given in Sect. 3.4.4 says that it converges if and only if for an arbitrary small  $\varepsilon > 0$ , we can find an  $X$  that yields

$$\left| \int_{A'}^A g'(x)e^{-wx}dx \right| < \varepsilon \quad (13.26)$$

with  $A'$ ,  $A > X$ . Therefore, our task is to show that the relation (13.26) holds for  $\operatorname{Re}(s) > \operatorname{Re}(s_0)$ .

Integration by parts gives us

$$\int_{A'}^A g'(x)e^{-wx}dx = -g(A')e^{-wA'} + g(A)e^{-wA} + w \int_{A'}^A g(x)e^{-wx}dx, \quad (13.27)$$

which results in

$$\left| \int_{A'}^A g'(x)e^{-wx}dx \right| < |g(A')|e^{-xA'} + |g(A)|e^{-xA} + |z| \int_{A'}^A |g(x)e^{-wx}|dx. \quad (13.28)$$

From (13.24) and from the hypothesis given in the theorem, it follows that for an arbitrary small  $\varepsilon' > 0$ , there exists a number  $X$  such that

$$|g(x)| < \varepsilon' \quad \text{when } x > X.$$

Thus, if  $A'$ ,  $A > X$  we have

$$|g(A')|, |g(A)| < \varepsilon'.$$

In addition, if

$$u \equiv \operatorname{Re}(w) > 0,$$

then the relation (13.28) becomes

$$\begin{aligned} \left| \int_{A'}^A g'(x)e^{-wx}dx \right| &< \varepsilon' \left[ e^{-uA'} + e^{-uA} + \frac{|w|}{u} (e^{-uA'} - e^{-uA}) \right] \\ &< \varepsilon' \left( 2 + \frac{|w|}{u} \right). \end{aligned} \quad (13.29)$$

Observe that the quantity in parentheses in (13.29) is finite for any fixed value of  $w$  with  $u > 0$ . Therefore, by making  $\varepsilon'$  small enough, the quantity

$$\varepsilon \equiv \varepsilon' \left( 2 + \frac{|w|}{u} \right) \quad (13.30)$$

becomes arbitrarily small; this can be the  $\varepsilon$  in the relation (13.26). Consequently, the relation (13.26) holds for any  $u > 0$ , or equivalently, for any

$$u = \operatorname{Re}(w) = \operatorname{Re}(s) - \operatorname{Re}(s_0) > 0.$$

This completes the proof of the theorem. (Note that if  $u = 0$ , the quantity in parentheses in (13.29) diverges, and if  $u < 0$ , the inequality (13.29) itself does not hold.) ♣

*Remark.* The theorem is inconclusive for the convergence property on the line  $\operatorname{Re}(s) = \operatorname{Re}(s_0)$  depicted on the complex  $s$ -plane. Note that we do not get convergence when  $\operatorname{Re}(s) = \sigma_0$ . This means that even though the integral converges at a point on the line of  $\operatorname{Re}(s) = \sigma_0$ , it does not necessarily converge all along the same line. A simple example is given by

$$f(x) = \begin{cases} 0, & 0 \leq x < 1, \\ \frac{1}{x}, & x \geq 1. \end{cases}$$

The Laplace integral

$$\int_0^\infty f(x) e^{-s_0 x} dx = \int_1^\infty \frac{e^{-i\omega_0 x}}{x} dx = \int_1^\infty \frac{\cos \omega_0 x}{x} dx - i \int_1^\infty \frac{\sin \omega_0 x}{x} dx$$

converges for  $s_0 = 0 + i\omega_0$  with  $\omega_0 \neq 0$ , but diverges at  $s_0 = 0$ .

### 13.3.5 Uniform Convergence for General Cases

A sufficient condition for uniform convergence is obtained in a similar way as in Sect. 13.3.4, although it is not the same as that for ordinary convergence. The difference is due to the fact that in the proof above,  $\varepsilon$  defined by (13.30) is dependent on  $s$  through  $|w| = |s - s_0|$ . In order to get the range of uniform convergence, we need a certain infinitesimal factor that can be taken independently of  $s$ .

To derive such a factor, let  $\theta$  be the angle of  $w = s - s_0$ , and observe that  $u = \operatorname{Re}(w)$  satisfies the relation

$$\frac{|w|}{u} = \left| \frac{1}{\cos \theta} \right|,$$

when  $u > 0$ . If  $\theta$  is restricted to the range

$$|\theta| < \frac{\pi}{2}, \quad (13.31)$$

we can find an angle  $\theta'$  that satisfies

$$|\theta| \leq \theta' < \frac{\pi}{2},$$

or equivalently,

$$\frac{|w|}{u} = \frac{1}{\cos \theta} \leq \frac{1}{\cos \theta'}.$$

Inserting this into (13.30), we have

$$\varepsilon = \varepsilon' \left( 2 + \frac{|w|}{u} \right) \leq \varepsilon' \left( 2 + \frac{1}{\cos \theta'} \right) \equiv \varepsilon'',$$

where the quantity  $\varepsilon''$  is independent of  $s$  and becomes arbitrarily small by making  $\varepsilon'$  small enough. This is true as far as condition (13.31) is satisfied; in this context, (13.31) represents the region of uniform convergence of the Laplace integral. Rewriting  $\theta$  by  $\arg(s - s_0)$ , we arrive at the following theorem:

♠ **Theorem** (= A sufficient condition for uniform convergence of the Laplace integrals for general cases):

If the improper integral

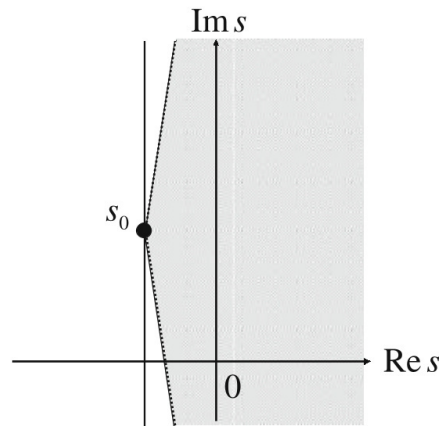
$$\int_0^\infty f(x)e^{-sx} dx$$

converges for  $s = s_0$ , then it converges uniformly to  $F(s) = L[f(x)]$  for

$$|\arg(s - s_0)| \leq \theta' < \frac{\pi}{2}.$$

(See also Fig. 13.8.)

Here, the  $\theta'$  shows the **closedness** of the converging region. The  $\theta'$  can be arbitrarily close to but not equal  $\pi/2$ .



**Fig. 13.8.** The region of uniform convergence for the Laplace integral that converges for  $s = s_0$

### 13.3.6 Distinction Between Exponential-Order Cases and General Cases

We have thus far presented four convergence theorems in connection with Laplace integrals, where the former two are associated with functions of exponential order and the latter two are relevant to more general functions. The theorems for the two cases are similar to the extent that they all identify a half-plane of convergence for the Laplace integral. Moreover, the general cases that we have considered cover a wide class of functions that includes exponential-order functions as a special case. At first glance, these remarks appear to imply that each of the former two theorems for exponential-order cases is a special case of each of the latter for general cases, but, *this is not true at all*. Below we give the reasons for this not being so.

First, the theorem for ordinary convergence in the exponential-order case is intrinsically different from that in the general case. Observe that the former theorem not only tells us that the Laplace integral converges in a half-plane; it also gives a specific number (i.e.,  $\alpha_0$ ) for the abscissa of a left-hand boundary of such a half-plane. (Of course  $\sigma_c \leq \alpha_0$  since it gives a *sufficient* condition for convergence.) In contrast, the latter theorem merely states convergence to the right of any point at which we already know that the integral converges; it gives no information about a boundary of the region of convergence.

Second, the regions of uniform convergence are specified in a different manner for the two cases. Whereas the theorem for general cases tells us only that the Laplace integral converges uniformly in an angular sector of the right half-plane, the theorem for exponential-order cases indicates uniform convergence in a less restricted region, namely, a half-plane.

In short, the theorems for the two cases are essentially different. As well, it should be emphasized again that all the four theorems provide *sufficient*

conditions for convergence of the Laplace integrals—not *necessary* or *necessary and sufficient* conditions.

### 13.3.7 Analytic Property of Laplace Transforms

An important consequence of uniform convergence of the Laplace integral is the fact that the corresponding Laplace transform,

$$F(s) = \int_0^{\infty} f(x)e^{-sx}dx, \quad (13.32)$$

is an **analytic function** on the complex  $s$ -plane. We know that if  $F(s)$  is analytic, it will exist outside the range of convergence of its integral representation, which can be uniquely determined by **analytic continuation**. From a practical viewpoint, the **analyticity** of  $F(s)$  plays a crucial role in evaluating the Laplace transform of a given function, since we can use it to treat the complex variable  $s$  as if it were real (see Sect. 13.1.3). We close this section by proving the analyticity of  $F(s)$ .

♠ **Theorem:**

The Laplace transform  $F(s)$  is analytic in the region of uniform convergence of the corresponding Laplace integral (13.32).

*Proof* We first recall that for  $F(s)$ , there is a region of uniform convergence in the  $s$ -plane and then we perform a contour integration with respect to  $s$  over an arbitrary simple closed path  $C$  in this region. Owing to the uniform convergence property, the order of integration may be inverted so that we have

$$\oint_C F(s)ds = \int_0^{\infty} f(x) \left( \oint_C e^{-sx}ds \right) dx = 0.$$

which gives us zero because **Cauchy's integral formula** means that

$$\oint_C e^{-sx}ds = 0.$$

Since the path of  $C$  is arbitrary in the region of uniform convergence, Morera's theorem establishes that  $F(s)$  is analytic inside the region of uniform convergence of its corresponding Laplace integral. ♣

## 13.4 Inverse Laplace Transform

### 13.4.1 The Two-Sided Laplace Transform

This section describes the **inverse Laplace transformation**. Intuitively understood, the inverse Laplace transform  $L^{-1}[F(s)]$  of a function  $F(s)$  is a

function  $f(x)$  whose Laplace transform is  $F(s)$ . Nevertheless, actual operations represented by the operator  $L^{-1}$  take some time to develop. To set to the explicit formula for manipulating the inverse transformation, we first introduce another kind of Laplace transform:

♠ **Two-sided Laplace transform:**

If the improper integral

$$\int_{-\infty}^{\infty} f(x)e^{-sx}dx \quad (13.33)$$

exists, it is called the **two-sided Laplace transform** (or **bilateral Laplace transform**), designated by  $\mathcal{F}(s) = \mathcal{L}[f(x)]$ .

It is easy to determine the region of convergence of such an integral. Observe that

$$\mathcal{L}[f(s)] = \int_{-\infty}^0 f(x)e^{-sx}dx + \int_0^{\infty} f(x)e^{-sx}dx. \quad (13.34)$$

The second integral is an ordinary Laplace integral so that it converges on a half-plane right to a fixed point denoted by  $x = \sigma_{c1}$ . By the change of variable  $x = -u$  the first integral becomes

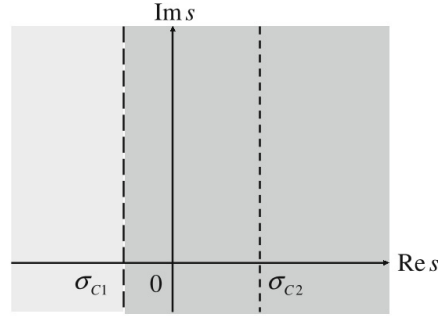
$$\int_{-\infty}^0 f(x)e^{-sx}dx = \int_0^{\infty} f(-u)e^{su}du.$$

Here, the latter integral is also an ordinary Laplace integral, although  $s$  has been replaced by  $-s$ . Hence, its region of convergence is a half-plane *left* to a point, say  $x = \sigma_{c2}$ . As a result, the common part of the two half-planes,  $\sigma_{c1} < \operatorname{Re}(s) < \sigma_{c2}$ , forms the region of convergence of the integral (13.34) as depicted in Fig. 13.9.

**Remark.** Typically, the range of convergence of (13.34) forms a vertical strip with a finite interval, but may be a right half plane, a left half-plane, the whole  $s$ -plane, a single point, or fail to exist.

**Example** We show that the function  $1/(s^2 + s)$  can be expressed as a two-sided Laplace integral. We readily see that

$$\frac{1}{s(s+1)} = \frac{1}{s} - \frac{1}{s+1}.$$



**Fig. 13.9.** Overlapping region: the region of convergence of the two-sided Laplace integral

We know that

$$\frac{1}{s+1} = \int_0^{\infty} e^{-x} e^{-sx} dx \quad \text{for } \sigma > -1 \quad (13.35)$$

and

$$\frac{1}{s} = \int_0^{\infty} e^{-sx} dx = \int_{-\infty}^0 e^{sx} dx \quad \text{for } \sigma > 0$$

the latter and that can be rewritten as

$$\frac{1}{s} = \int_{-\infty}^0 e^{-sx} (-1) dx \quad \text{for } \sigma < 0. \quad (13.36)$$

From (13.35) and (13.36), we obtain

$$\frac{1}{s(s+1)} = \frac{1}{s} - \frac{1}{s+1} = \int_{-\infty}^{\infty} f(x) e^{-sx} dx,$$

where

$$f(x) = \begin{cases} -e^{-x}, & 0 < x < \infty, \\ -1, & -\infty < x < 0, \end{cases}$$

which means that  $1/(s^2 + s) = \mathcal{L}[f(x)]$ . The interval of convergence is seen to be  $-1 < \sigma < 0$ . example

### 13.4.2 Inverse of the Two-Sided Laplace Transform

Having introduced the two-sided Laplace transform, we are ready to undertake the inverse Laplace transformation. We first observe that the two-sided Laplace transform

$$\mathcal{F}(s) = \int_{-\infty}^{\infty} f(x) e^{-sx} dx \quad (13.37)$$

is identical with the **Fourier transform**

$$\mathcal{F}(\sigma + i\omega) = \int_{-\infty}^{\infty} f(x)e^{-\sigma x}e^{-i\omega x}dx$$

if we regard the real number  $\sigma$  as fixed. We use the **inverse Fourier transformation** to yield

$$f(x)e^{-\sigma x} = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathcal{F}(\sigma + i\omega)e^{i\omega x}d\omega,$$

or equivalently,

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathcal{F}(\sigma + i\omega)e^{\sigma x}e^{i\omega x}d\omega. \quad (13.38)$$

We then replace  $\sigma + i\omega$  by  $s$ , keeping in mind that  $s$  should lie on the vertical line with the abscissa  $\operatorname{Re}(s) = \sigma$ . Then the integral (13.38) can be regarded as a contour integral along the vertical line  $\operatorname{Re}(s) = \sigma$ . On this contour,

$$ds = i d\omega,$$

so the integral (13.38) becomes

$$f(x) = \frac{1}{2\pi i} \int_{\sigma-i\infty}^{\sigma+i\infty} \mathcal{F}(s)e^{sx}ds \quad (\operatorname{Re}(s) = \sigma \text{ is fixed}). \quad (13.39)$$

This result provides a clue for evaluating the explicit form of  $f(x)$  from its two-sided Laplace transform  $\mathcal{F}(s)$ .

The result (13.39) is not yet satisfying. We should recall that  $f(x)$  is not determined uniquely by  $\mathcal{F}(s)$  through (13.39) unless the location of the point  $x = \sigma$  is specified (see Exercise 3 in Sect. 13.4). If we know in advance that  $\sigma$  lies in the region of convergence of the two-sided integral given by (13.37), i.e., the strip of convergence,  $f(x)$  is uniquely determined by (13.39). However, if  $\sigma$  used in (13.39) is set outside this strip, the integral of (13.39) is altered quantitatively because the integration contour passes over one or more singular points of  $\mathcal{F}(s)$ . Thus for us to be able to use equation (13.39), we must know the region of convergence of the Laplace integral of  $f(x)$  *before* we can fix the real number  $\sigma$ . If only  $\mathcal{F}(s)$  is given, we will not be able to locate this region, and not be able to obtain  $f(x)$  because we will not know where to put  $\sigma$ . These caveats lead to the following theorem:

♠ **Theorem:**

The inverse of the two-sided Laplace transform

$$f(x) = \frac{1}{2\pi i} \int_{\sigma-i\infty}^{\sigma+i\infty} \mathcal{F}(s) e^{-sx} ds \quad (\operatorname{Re}(s) = \sigma \text{ is fixed})$$

determines  $f(x)$  uniquely only if we know where  $\sigma$  should be located.

### 13.4.3 Inverse of the One-Sided Laplace Transform

Let us develop the theory that correspond to the above for the one-sided transform. We compare the two-sided transform  $\mathcal{L}[f(x)]$  and its one-sided counterpart  $L[f(x)]$ , where  $f(x)$  is the same function in both cases and is defined for all  $x$ . From the definitions of the one- and two-sided transforms, it is evident that

$$F(s) = L[f(x)] = \mathcal{L}[f(x)\theta(x)],$$

where  $\theta(x)$  is the step function. This implies that

$$f(x)\theta(x) = \frac{1}{2\pi i} \int_{\sigma-i\infty}^{\sigma+i\infty} F(s) e^{sx} ds \quad (\sigma \text{ is fixed}). \quad (13.40)$$

Here,  $\sigma$  must be to the right of all the singularities of  $F(s)$  in order for the integral in (13.40) to converge. As a consequence, we have arrived at the following theorem:

♠ **The inverse Laplace transformation:**

If the function  $F(s)$  defined by

$$F(s) = \int_0^\infty e^{-sx} f(x) dx$$

is analytic for  $\operatorname{Re}(s) > \sigma_c$ , then  $f(x)$  for  $x > 0$  is uniquely determined by

$$f(x) = \lim_{\omega \rightarrow \infty} \frac{1}{2\pi i} \int_{\sigma-i\omega}^{\sigma+i\omega} e^{sx} F(s) ds,$$

where  $\sigma$  is arbitrary for all  $\sigma > \sigma_c$ .

### 13.4.4 Useful Formula for Inverse Laplace Transformation

In contrast to the situation with the inverse Fourier transformation, the use of the inverse Laplace transformation formula is less convenient. This is primarily because the calculation of the complex integral

$$\int_{\sigma-i\infty}^{\sigma+i\infty} e^{sx} F(s) ds$$

can be rather complicated. In this subsection, we present a simple and natural method of computing integrals of this form that is based on the **residue theorem**.

Suppose that  $F(s)$  is analytic on the domain  $\operatorname{Re}(s) > \sigma_c$ . We wish to compute

$$f(x) = \lim_{\omega \rightarrow \infty} \frac{1}{2\pi i} \int_{\sigma-i\omega}^{\sigma+i\omega} e^{sx} F(s) ds, \quad x > 0.$$

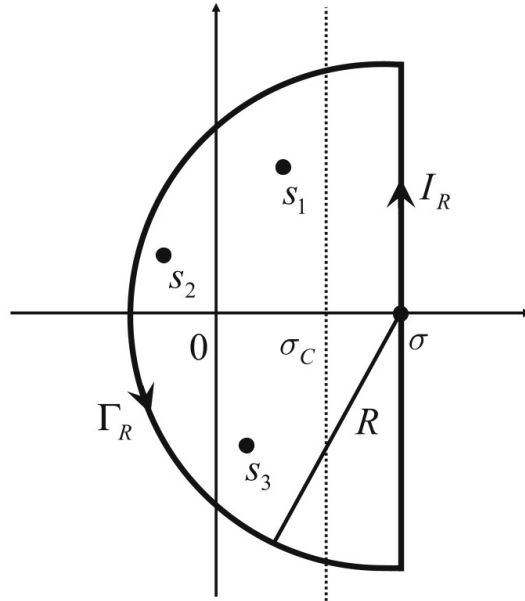
No general method for doing this exists, but it is possible to evaluate this integral under certain conditions on  $F(s)$ . Suppose that  $F(s)$  is analytic on the entire complex plane, except at a finite number of singularities  $s_1, s_2, \dots, s_n$  satisfying

$$\operatorname{Re}(s_j) < \sigma_c, \quad j = 1, 2, \dots, n.$$

Figure 13.10 is a sketch for this situation. Let  $\sigma > \sigma_c$  and let  $R > 0$  be a real number sufficiently large that the left half-circle  $C_R$  with center  $s = \sigma$  and radius  $R$  encloses all the points  $s_1, s_2, \dots, s_n$ . Divide  $C_R$  into the two segments:

$$I_R = \{s \in C : s = \sigma + i\omega, -R < \omega < R\},$$

$$\Gamma_R = \{s \in C : |s - \sigma| = R, \operatorname{Re}(s) \leq \sigma\}.$$



**Fig. 13.10.** A finite number of singularities  $s_j$  of  $F(s)$  enclosed by the *left half-circle*  $C_R$  composed of  $\Gamma_R$  and  $I_R$

By the residue theorem,

$$\oint_{C_R} e^{sx} F(s) ds = \sum_{j=1}^n \operatorname{Res} [e^{sx} F(s); s_j].$$

The right-hand side is independent of  $R$ , if  $R$  is sufficiently large. From  $C_R = \Gamma_R \cup I_R$ , it follows that

$$\oint_{C_R} e^{sx} F(s) ds = \int_{\Gamma_R} e^{sx} F(s) ds + \int_{I_R} e^{sx} F(s) ds.$$

Clearly,

$$\lim_{M \rightarrow \infty} \int_{\sigma-iM}^{\sigma+iM} e^{sx} F(s) ds = \lim_{R \rightarrow \infty} \int_{I_R} e^{sx} F(s) ds.$$

Therefore if, by chance, we have

$$\lim_{R \rightarrow \infty} \int_{\Gamma_R} e^{sx} F(s) ds = 0, \quad (13.41)$$

then we obtain

$$f(x) = \lim_{M \rightarrow \infty} \frac{1}{2\pi i} \int_{\sigma-iM}^{\sigma+iM} e^{sx} F(s) ds = \sum_{j=1}^n \operatorname{Res} [e^{sx} F(s); s_j].$$

Unfortunately, condition (13.41) does not hold for every  $F$ . The next theorem presents a sufficient condition on  $F$  under which (13.41) holds true.

♠ **Theorem:**

Let  $F$  be an analytic function on the complex plane except at a finite number of points (if they exist) and let  $\Gamma_R$  be as above. If

$$\lim_{R \rightarrow \infty} \max_{s \in \Gamma_R} |F(s)| = 0,$$

then

$$\lim_{R \rightarrow \infty} \int_{\Gamma_R} e^{sx} F(s) ds = 0$$

holds for every  $x > 0$ .

*Proof* This theorem is a reinterpretation of **Jordan's lemma** given in Sect. 9.2.4. ♣

An immediate consequence of this theorem is the following:

♠ **Theorem:**

Let  $F$  be an analytic function on the complex plane except at a finite number of points  $s_1, s_2, \dots, s_n$ , satisfying  $\operatorname{Re}(s_j) < \sigma$  for all  $j$ . If

$$\lim_{R \rightarrow \infty} \max_{s \in C_R} |F(s)| = 0, \quad (13.42)$$

then the inverse Laplace transform of  $F(s)$  is given by

$$f(x) = \sum_{j=1}^n \operatorname{Res} [e^{sx} F(s); s_j]. \quad (13.43)$$

### 13.4.5 Evaluating Inverse Transformations

Below are several examples of actual evaluations of inverse Laplace transforms via the residue formula (13.43).

*Example 1.* Assume a complex-valued function

$$F(s) = \frac{1}{s^2 - 3s + 2}$$

that has two simple poles  $s_1 = 1$  and  $s_2 = 2$ . We thus choose  $\sigma = 3$  and set

$$C_R = \{s : |s - 3| = R, \operatorname{Re}(s) \leq 3\}$$

in order to make use of equation (13.43). Before doing so, we must check that condition (13.42) is satisfied. Observe that

$$\max_{s \in C_R} |F(s)| = \max_{s \in C_R} \left| \frac{1}{(s-1)(s-2)} \right|.$$

If we let  $R = |s - 3|$  go to infinity, then  $|s - 1|$  and  $|s - 2|$  will also converge to infinity, so that

$$\lim_{R \rightarrow \infty} \max_{s \in C_R} \frac{1}{|(s-1)(s-2)|} = 0.$$

Thus (13.43) provides the desired result:

$$\begin{aligned} f(x) &= \operatorname{Res} \left[ \frac{e^{sx}}{(s-1)(s-2)}; s=1 \right] + \operatorname{Res} \left[ \frac{e^{sx}}{(s-1)(s-2)}; s=2 \right] \\ &= \frac{e^{sx}}{s-2} \Big|_{s=1} + \frac{e^{sx}}{s-1} \Big|_{s=2} \\ &= \frac{e^x}{1-2} + \frac{e^{2x}}{2-1} = -e^x + e^{2x}. \quad \clubsuit \end{aligned}$$

*Remark.* The above example can be solved more easily by rewriting  $F$  using partial fractions  $F(s) = 1/(s-2) - 1/(s-1)$ , followed by applying known equations to get

$$f(x) = L^{-1}[F(s)] = L^{-1}\left[\frac{1}{s-2}\right] - L^{-1}\left[\frac{1}{s-1}\right] = e^{2x} - e^x.$$

*Example 2.* It should be cautioned that equation (13.43) is valid only when the condition (13.42) is satisfied. As a negative example, let us consider the step function

$$\theta(x) = \begin{cases} 1, & x > c, \\ 0, & x < c, \end{cases}$$

with  $c > 0$ , whose Laplace transform reads

$$F(s) = L[\theta(x)] = \frac{e^{-cs}}{s} \quad (s > 0),$$

We would like to derive  $\theta(x)$  from  $F(s)$  through the inverse transformation given by

$$f(x) = L^{-1}[F(s)] = \lim_{M \rightarrow \infty} \frac{1}{2\pi i} \int_{\sigma-iM}^{\sigma+iM} \frac{e^{sx} e^{-cs}}{s} ds \quad (0 < x \neq c).$$

However, we cannot use (13.43) to obtain  $f(x)$ , since the function  $e^{-cs}/s$  does not satisfy condition (13.42). In fact, if we set  $s = \sigma - R$ , then

$$\max_{s \in C_R} \left| \frac{e^{-cs}}{s} \right| \geq \frac{e^{cR} e^{-c\sigma}}{|\sigma - R|} \rightarrow \infty \quad (R \rightarrow \infty)$$

since  $c > 0$ .

*Remark.* If we were to use (13.43) in Example 2, we would obtain a wrong result. The function  $e^{-cs}/s$  has a single simple pole at  $s = 0$ , so

$$\text{Res} \left[ \frac{e^{sx} e^{-cs}}{s}; s = 0 \right] = 1$$

for each value of  $x$ . This is, of course, not the step function  $\theta(x)$ .

*Example 3.* Next we consider the inverse Laplace transformation  $L^{-1}[F(s)]$  of the function

$$F(s) = \frac{1}{s+a} \quad (a > 0).$$

The  $F(s)$  has a first-order pole at  $s = -a$ . The residue of  $F(s)e^{sx}$  at  $s = -a$  reads

$$\text{Res}[F(s)e^{sx}, -a] = \lim_{s \rightarrow -a} (s + a)F(s)e^{sx} = \lim_{s \rightarrow -a} e^{sx} = e^{-ax}.$$

Hence, we have

$$f(x) = L^{-1}[F(s)] = e^{-ax} \quad (x > 0).$$

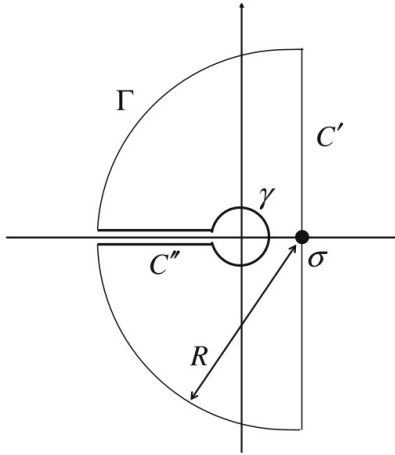
### 13.4.6 Inverse Transform of Multivalued Functions

Some caution must be taken when considering the inverse Laplace transform of **multivalued functions**. As an example, we consider a multivalued function  $F(s) = 1/\sqrt{s}$ , and examine its inverse transform given by

$$f(x) = \frac{1}{2\pi i} \int_C \frac{e^{sx}}{\sqrt{s}} ds, \quad (13.44)$$

where the symbol  $\sqrt{s}$  represents values of the original double-valued function  $s^{1/2}$  in the same sheet of the **Riemann surface**. The function  $1/\sqrt{s}$  has a **branch point** at  $s = 0$ , so among many choices we set its **branch cut** at  $(-\infty, 0]$ .

Since the function  $1/\sqrt{s}$  approaches zero as  $|s| \rightarrow \infty$ , **Jordan's lemma** is applicable. Nevertheless, the problem becomes rather complicated owing to the presence of the branch cut. To perform the integration of (13.44), we close the path  $\Gamma$  by a circle to the left, bypassing the branch cut in the manner shown in Fig. 13.11. No singularities are enclosed by the closed curve consisting of  $C' + \Gamma + \gamma + C''$ , in which  $C'$  is the vertical line,  $C''$  is the pair of parallel



**Fig. 13.11.** Closed loop employed in evaluating the integral (13.44)

horizontal segments,  $\gamma$  is the small circle of radius  $\delta$ , and  $\Gamma$  is a semicircle from which the infinitesimal gap at the branch cut has been omitted. Hence, we have

$$\int_{C'} \frac{e^{sx}}{\sqrt{s}} ds + \int_{\Gamma} \frac{e^{sx}}{\sqrt{s}} ds + \int_{\gamma} \frac{e^{sx}}{\sqrt{s}} ds + \int_{C''} \frac{e^{sx}}{\sqrt{s}} ds = 0.$$

In the limit  $R \rightarrow \infty$ , the integral over  $\Gamma$  vanishes and the path  $C'$  reduces to  $C$  as given in (13.44), which implies that

$$f(x) = \lim_{R \rightarrow \infty} \frac{1}{2\pi i} \int_C \frac{e^{sx}}{\sqrt{s}} ds = \lim_{R \rightarrow \infty} \frac{-1}{2\pi i} \int_{C''+\gamma} \frac{e^{sx}}{\sqrt{s}} ds. \quad (13.45)$$

Thus our remaining task is to evaluate the last term in (13.45).

Recall that  $\sqrt{s}$  is double-valued so that it is discontinuous across the branch cut. Consequently, on the parallel segments  $C''$ ,

$$\sqrt{s} = i\sqrt{\rho} \quad \text{and} \quad -i\sqrt{\rho} \quad \text{with } \rho = |s|,$$

above and below the branch cut, respectively. On the small circle  $\gamma$ ,

$$\sqrt{s} = \sqrt{\delta} e^{i\phi/2},$$

where  $-\pi < \phi \leq \pi$ , and  $ds = -d\rho$  on each of the straight lines. Thus, we have

$$\begin{aligned} \int_{C''+\gamma} \frac{e^{sx}}{\sqrt{s}} ds &= -i \int_{R-\sigma_1}^{\delta} \frac{e^{-\rho x}}{\sqrt{\rho}} (-d\rho) + i \int_{\delta}^{R-\sigma_1} \frac{e^{-\rho x}}{\sqrt{\rho}} (-d\rho) \\ &\quad + i \frac{\delta}{\sqrt{\delta}} \int_{\pi}^{-\pi} \frac{e^{(\delta \cos \phi)x} e^{i(\delta \sin \phi)x} e^{i\phi}}{e^{i\phi/2}} d\phi. \end{aligned}$$

Let  $\delta$  go to zero and  $R$  approach infinity; then, the first two integrals on the right-hand side combine into a single integral. The last integral on the right-hand side approaches zero. As a result, we have

$$\lim_{\delta \rightarrow 0} \lim_{R \rightarrow \infty} \int_{C''+\gamma} \frac{e^{sx}}{\sqrt{s}} ds = -2i \int_0^{\infty} \frac{e^{-\rho x}}{\sqrt{\rho}} d\rho,$$

which implies that

$$f(x) = \frac{1}{\pi} \int_0^{\infty} \frac{e^{-\rho x}}{\sqrt{\rho}} d\rho.$$

By substituting  $\rho x = u^2$ , the right-hand side becomes

$$\frac{1}{\pi} \int_0^{\infty} \frac{e^{-\rho x}}{\sqrt{\rho}} d\rho = \frac{2}{\pi\sqrt{x}} \int_0^{\infty} e^{-u^2} du = \frac{1}{\sqrt{\pi x}}.$$

Eventually, we obtain

$$f(x) = L^{-1} \left( \frac{1}{\sqrt{s}} \right) = \frac{1}{\sqrt{\pi x}},$$

which is consistent with the earlier result presented in (13.15).

## Exercises

1. Find (a)  $L^{-1}[5/(p+2)]$  and (b)  $L^{-1}[1/p^s]$  where  $s > 0$ .

**Solution:** (a) Recall that  $L[e^{ax}] = 1/(p-a)$ ; hence  $L^{-1}[1/(p-a)] = e^{ax}$ . It follows that

$$L^{-1} \left[ \frac{5}{p+2} \right] = 5L^{-1} \left[ \frac{1}{p+2} \right] = 5e^{-2x}.$$

(b) Recall that

$$L[x^k] = \int_0^\infty e^{-sx} x^k dx = \frac{\Gamma(k+1)}{p^{k+1}}.$$

From this we have

$$L \left[ \frac{x^k}{\Gamma(k+1)} \right] = \frac{1}{p^{k+1}},$$

so

$$L^{-1} \left[ \frac{1}{p^{k+1}} \right] = \frac{x^k}{\Gamma(k+1)}.$$

If we now let  $k+1 = s$ , then

$$L^{-1} \left[ \frac{1}{p^s} \right] = \frac{x^{s-1}}{\Gamma(s)}. \clubsuit$$

2. Solve the differential equation

$$f''(x) + f(x) = 1 \tag{13.46}$$

with the initial conditions

$$f(0) = f'(0) = 0$$

using the Laplace transformation.

**Solution:** Taking the Laplace transform of both sides of (13.46), we obtain

$$L[f''(x)] + L[f(x)] = L[1]. \quad (13.47)$$

Substituting the result

$$L[f''(x)] = s^2 L[f(x)] - s \cdot f(0) - f'(0) = s^2 L[f(x)]$$

and  $L[1] = 1/s$  into (13.47) yields

$$s^2 L[f(x)] + L[f(x)] = \frac{1}{s},$$

i.e.,

$$L[f(x)] = F(s) = \frac{1}{s(s^2 + 1)} = \frac{1}{s} - \frac{s}{s^2 + 1}.$$

Thus we see that

$$\begin{aligned} f(x) &= L^{-1}[F(s)] = L^{-1}\left[\frac{1}{s}\right] - L^{-1}\left[\frac{s}{s^2 + 1}\right] \\ &= \begin{cases} 1 - \cos x & \text{for } x \geq 0, \\ 0 & \text{for } x < 0, \end{cases} \end{aligned}$$

which is the solution of the initial value problem originally given by (13.47). ♣

**3.** Derive the two-sided Laplace transform of the following three functions:

$$f_a(x) = \begin{cases} e^{-2x} - e^{-x}, & x > 0, \\ 0, & x < 0, \end{cases} \quad f_b(x) = \begin{cases} e^{-2x}, & x > 0, \\ e^{-x}, & x < 0, \end{cases}$$

and

$$f_c(x) = \begin{cases} 0, & x > 0, \\ e^{-2x} - e^{-x}, & x < 0. \end{cases}$$

**Solution:** The two-sided Laplace transform read, respectively,

$$\mathcal{L}[f_a(x)] = \frac{1}{s+2} - \frac{1}{s+1} \quad \text{for } \sigma > -1,$$

$$\mathcal{L}[f_b(x)] = \frac{1}{s+2} - \frac{1}{s+1} \quad \text{for } -2 < \sigma < -1,$$

$$\mathcal{L}[f_c(x)] = \frac{1}{s+2} - \frac{1}{s+1} \quad \text{for } \sigma < -2.$$

Clearly, all the  $s$  functions are the same and may be labeled  $\mathcal{F}(s)$  (although the region of convergence is different). This implies that the inverse of a two-sided transform is uniquely determined only after the location of  $\sigma$  is fixed. ♣

## 13.5 Applications in Physics and Engineering

### 13.5.1 Electric Circuits I

The most familiar applications of Laplace transformations in the physical sciences are encountered in analyses of electric circuits. Consider the RC circuit depicted in Fig. 13.12. The electric charge  $q(t)$  deposited in the condenser with capacitance  $C$  is governed by the equation

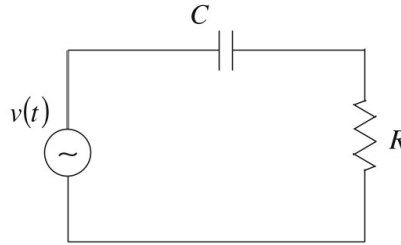
$$R \frac{dq(t)}{dt} + \frac{q(t)}{C} = v(t), \quad q(t=0) = 0, \quad (13.48)$$

where  $R$  is a resistance and  $v(t)$  is the external voltage. We set a rectangular voltage defined by (see Fig. 13.13)

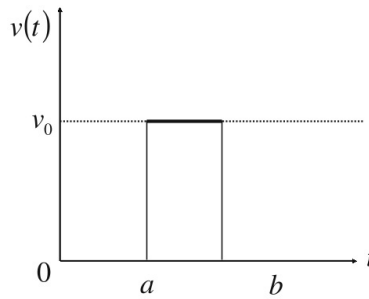
$$v(t) = v_0 \times [\theta(t-a) - \theta(t-b)] \quad (a < b) \quad (13.49)$$

with the step function

$$\theta(t-a) = \begin{cases} 0, & t < a, \\ 1, & t \geq a. \end{cases}$$



**Fig. 13.12.** Diagram of an RC circuit



**Fig. 13.13.** The time dependence of a rectangular voltage applied to the RC circuit

We now want to solve the differential equation (13.48) with respect to  $q(t)$ . To do this, we apply the Laplace transform to both sides of (13.48) and make use of the symbol  $Q(s) \equiv L[q(t)]$ . Straightforward calculation yields

$$sQ(s) + \frac{Q(s)}{\tau} = \frac{v_0}{R} \left( \frac{e^{-as}}{s} - \frac{e^{-bs}}{s} \right),$$

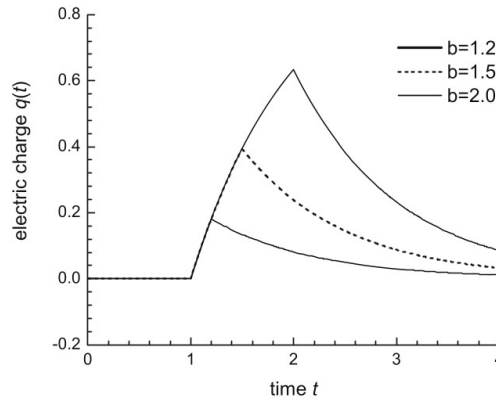
where  $\tau \equiv RC$  is called a **damping time constant**. Hence, we have

$$\begin{aligned} Q(s) &= \frac{v_0}{R} \frac{1}{s + \tau^{-1}} \left( \frac{e^{-as}}{s} - \frac{e^{-bs}}{s} \right) \\ &= Cv_0 \left( \frac{1}{s} - \frac{1}{s + \tau^{-1}} \right) (e^{-as} - e^{-bs}) \\ &= Cv_0 \left( \frac{e^{-as}}{s} - \frac{e^{-bs}}{s} - \frac{e^{-as}}{s + \tau^{-1}} + \frac{e^{-bs}}{s + \tau^{-1}} \right). \end{aligned}$$

Then we use the inverse transform to obtain

$$\begin{aligned} q(t) &= L^{-1}[Q(s)] \\ &= Cv_0 \left[ \theta(t-a) - \theta(t-b) - e^{-(t-a)/\tau} \theta(t-a) + e^{-(t-b)/\tau} \theta(t-b) \right], \\ &= \begin{cases} 0 & t < a, \\ cv_0 [1 - e^{-(t-a)/\tau}] & a < t < b, \\ cv_0 [e^{b/\tau} - e^{a/\tau}] e^{-t/\tau} & t > b. \end{cases} \end{aligned}$$

The explicit time-dependence of the charge  $q(t)$  given by (13.50) is illustrated in Fig. 13.14, in which various separations  $b-a$  are taken.



**Fig. 13.14.** Time dependence of the electric charge  $q(t)$  described by (13.50), which is accumulated in the condenser in the RC circuit. The parameter  $a$  introduced in (13.49) is fixed at  $a = 1.0$

### 13.5.2 Electric Circuits II

Next, in order to illustrate the use of **convolution integrals** in applications of Laplace transforms, we solve the previous equation (13.48) with respect to the current  $i(t)$  instead of charge  $q(t)$ . We consider the differential equation

$$Ri(t) + \frac{1}{C} \int_0^t i(u) du = v(t), \quad (13.50)$$

with the rectangular voltage (13.49). The integral term on the left-hand side in (13.50) is rewritten as a convolution integral:

$$\int_0^t i(u) du = \int_0^t i(u) \theta(t-u) du = \theta(t) * i(t), \quad (13.51)$$

whose Laplace transform reads

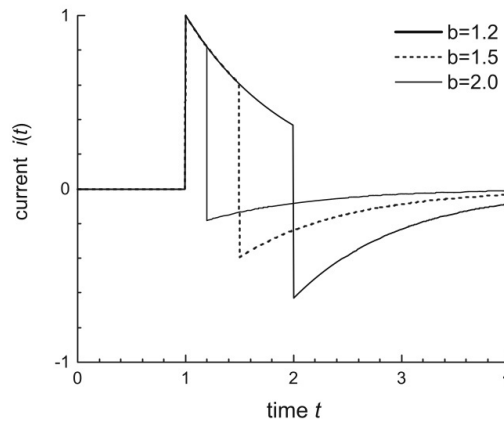
$$L[\theta(t) * i(t)] = L[\theta(t)] \cdot L[i(t)] = \frac{1}{s} I(s).$$

Hence, applying the Laplace transformation of both sides of (13.50) yields

$$RI(s) + \frac{I(s)}{Cs} = \frac{v_0}{s} (e^{-as} - e^{-bs}), \quad (13.52)$$

which implies

$$I(s) = \frac{v_0}{R} \frac{e^{-as} - e^{-bs}}{s + \tau^{-1}} \quad (\tau = RC). \quad (13.53)$$



**Fig. 13.15.** Time dependence of the current  $i(t)$  in the RC circuit described by (13.54). The parameter  $a$  introduced in (13.49) is fixed at  $a = 1.0$

Using the inverse transformation, we finally set

$$\begin{aligned}
 i(t) &= L^{-1} [I(s)] \\
 &= \frac{v_0}{R} \left[ e^{-(t-a)/\tau} \theta(t-a) - e^{-(t-b)/\tau} \theta(t-b) \right] \\
 &= \begin{cases} 0 & t < a, \\ \frac{v_0}{R} e^{a/\tau} e^{-t/\tau} & a \leq t < b, \\ \frac{v_0}{R} \left( e^{a/\tau} - e^{b/\tau} \right) e^{-t/\tau} & t \geq b. \end{cases}
 \end{aligned}$$

Figure 13.15 illustrates the time dependence of the current  $i(t)$ .

## Wavelet Transformation

**Abstract** Similar to the Fourier and Laplace transforms, a wavelet transform is an integral transform of a function by using “wavelets.” A wavelet is a mathematical mold with a finite-length and fast-decaying oscillating waveform, which is used to divide a given function into different scale components. Wavelet transforms have certain advantages over conventional Fourier transforms, as they can reveal the nature of a function in the time and frequency domains simultaneously.

### 14.1 Continuous Wavelet Analyses

#### 14.1.1 Definition of Wavelet

This short chapter covers the minimum ground for understanding **wavelet analysis**. The concept of **wavelet** originates from the study of signal analysis, i.e., from the need in certain cases to analyze a signal in the time and frequency domains simultaneously. The crucial advantage of wavelet analyses is that they allow us to decompose complicated information contained in a signal into elementary functions associated with different time scales and different frequencies and to reconstruct it with high precision and efficiency. In the following discussions, we first determine what constitutes a wavelet and then describe how it is used in the transformation of a signal.

The primary question concerns the definition of a wavelet:

##### ♠ Wavelet:

A **wavelet** is a real-valued function  $\psi(t)$  having a localized waveform that satisfies the following criteria:

1. The integral of  $\psi(t)$  is zero:  $\int_{-\infty}^{\infty} \psi(t) dt = 0$ .
2. The square of  $\psi(t)$  integrates to unity:  $\int_{-\infty}^{\infty} \psi(t)^2 dt = 1$ .

- 3.** The Fourier transform  $\Psi(\omega)$  of  $\psi(t)$  satisfies the **admissibility condition** expressed by

$$C_\Psi \equiv \int_0^\infty \frac{|\Psi(\omega)|^2}{\omega} d\omega < \infty. \quad (14.1)$$

Here,  $C_\Psi$  is called the **admissibility constant**, whose value depends on the chosen wavelet.

We restrict our attention to real-valued wavelets, although it is possible to define complex-valued wavelets as well. Observe that condition **2** above says that  $\psi(t)$  has to deviate from zero at finite intervals of  $t$ . On the other hand, condition **1** tells us that any deviation above zero must be canceled out by a deviation below zero. Hence,  $\psi(t)$  must oscillate across the  $t$ -axis like a wave. The following are the most important two examples of wavelets:

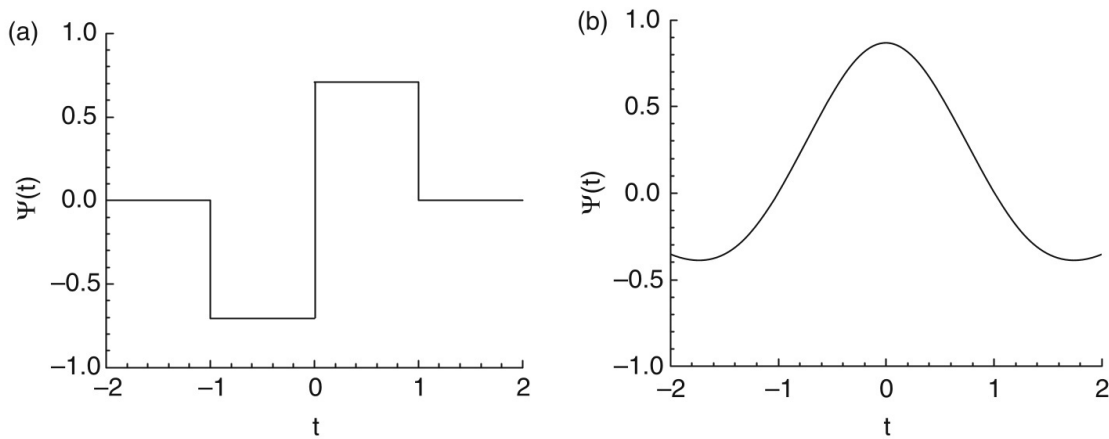
*Examples* **1.** The **Haar wavelet** (See Fig. 14.1a):

$$\psi(t) \equiv \begin{cases} -\frac{1}{\sqrt{2}}, & -1 < t \leq 0, \\ \frac{1}{\sqrt{2}}, & 0 < t \leq 1, \\ 0, & \text{otherwise.} \end{cases} \quad (14.2)$$

**2.** The **Mexican hat wavelet** (see Fig. 14.1b):

$$\psi(t) \equiv \frac{2 \left(1 - \frac{t^2}{\sigma^2}\right) e^{-t^2/(2\sigma^2)}}{\sqrt{3}\sigma\pi^{1/4}}. \quad (14.3)$$

To form the Mexican hat wavelet (14.3), we start with the Gaussian function with mean zero and variance  $\sigma^2$ :



**Fig. 14.1.** (a) The Haar wavelet given by (14.2). (b) The Mexican hat wavelet given by (14.3) with  $\sigma = 1$

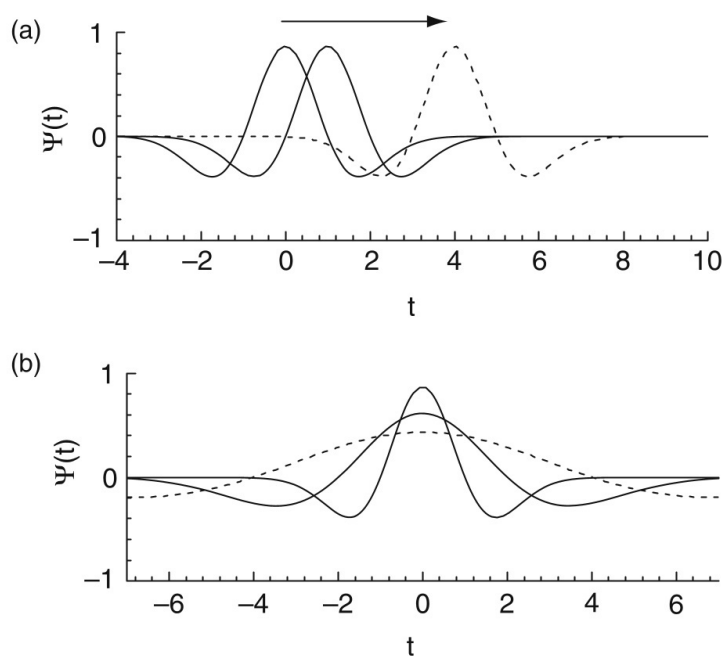
$$f(t) \equiv \frac{e^{-t^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}}.$$

If we take the negative of the second derivative of  $f(t)$  with normalization for satisfying condition **2**, we obtain the Mexican hat wavelet (14.3). In the meantime, we proceed with our argument on the basis of that wavelet by setting  $\sigma = 1$  and omitting the normalization constant for simplicity.

**Remark.** We know that all the derivatives of the Gaussian function may be used as wavelets. The most appropriate one many particular case depends on the application.

### 14.1.2 The Wavelet Transform

In mathematical terminology, the **wavelet transform** is known as a **convolution**; more precisely, it is a convolution of the wavelet function with a signal to be analyzed. In the convolution process, two parameters are involved that manipulate the function form of the wavelet. The first is the **dilatation parameter** denoted by  $a$ , which characterizes the dilation and contraction of the wavelet in the time domain (see Fig. 14.2a). For the Mexican hat wavelet, it is the distance between the center of the wavelet and its crossing of the time axis. The second is the **translation parameter  $b$ , which governs the movement of the wavelet along the time axis (see Fig. 14.2b). With this notation, shifted and dilated versions of a Mexican hat wavelet are expressed by**



**Fig. 14.2.** Translation (a) and dilatation of a wavelet (b)

$$\psi\left(\frac{t-b}{a}\right) = \left[1 - \left(\frac{t-b}{a}\right)^2\right] e^{-[(t-b)/a]^2/2}, \quad (14.4)$$

where we have set  $\sigma = 1$  in (14.3) and omitted the normalization factor for simplicity. We are now in a position to define the wavelet transform.

♠ **Wavelet transform:**

The **wavelet transform**  $T(a, b)$  of a continuous signal  $x(t)$  with respect to the wavelet  $\psi(t)$  is defined by

$$T(a, b) = w(a) \int_{-\infty}^{\infty} x(t) \psi\left(\frac{t-b}{a}\right) dt, \quad (14.5)$$

where  $w(a)$  is an appropriate **weight function**.

Typically,  $w(a)$  is set to  $1/\sqrt{a}$  because this choice yields

$$\int_{-\infty}^{\infty} \left[ \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{a}\right) \right]^2 dt = \int_{-\infty}^{\infty} \psi(u)^2 du = 1 \quad \text{with } u = \frac{t-b}{a},$$

i.e., the normalization condition for the square integral of  $\psi(t)$  remains invariant, which is why we use this value for the rest of this section.

The dilated and shifted wavelet is often written more compactly as

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{a}\right),$$

so that the transform integral may be written as

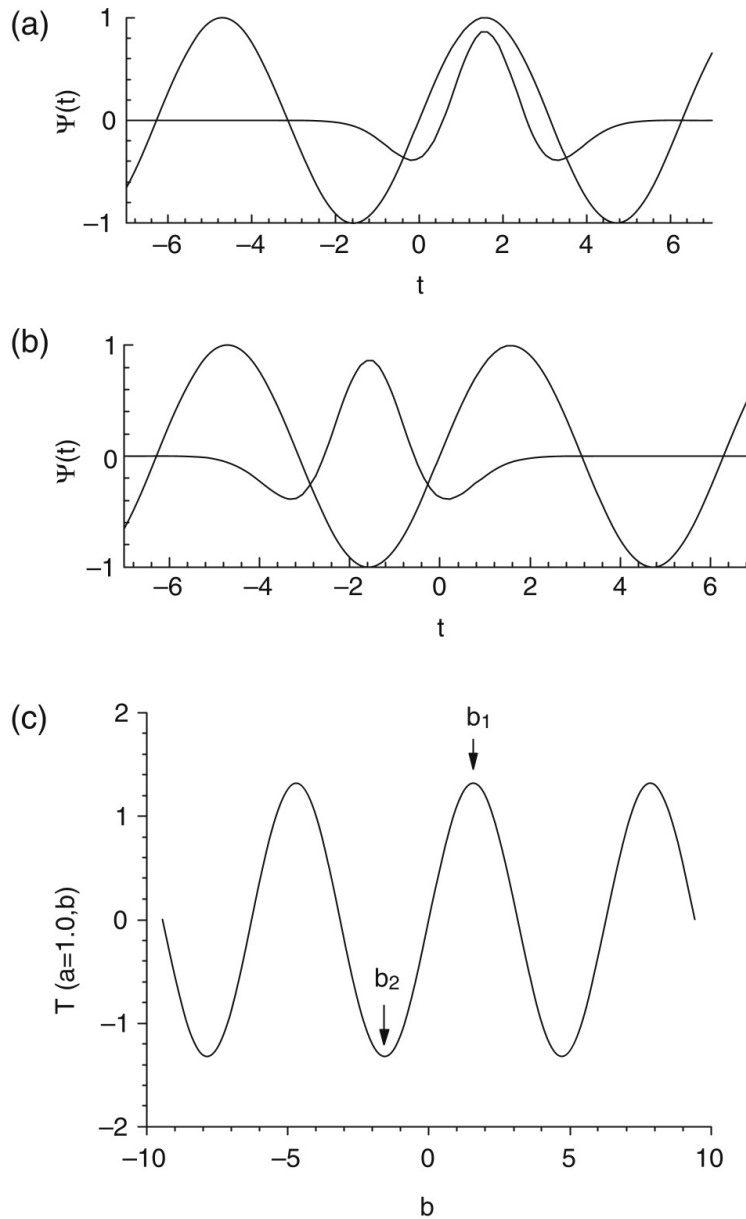
$$T(a, b) = \int_{-\infty}^{\infty} x(t) \psi_{a,b}(t) dt. \quad (14.6)$$

From here on, we use this notation and refer to  $\psi_{a,b}(t)$  simply as the wavelet.

### 14.1.3 Correlation Between Wavelet and Signal

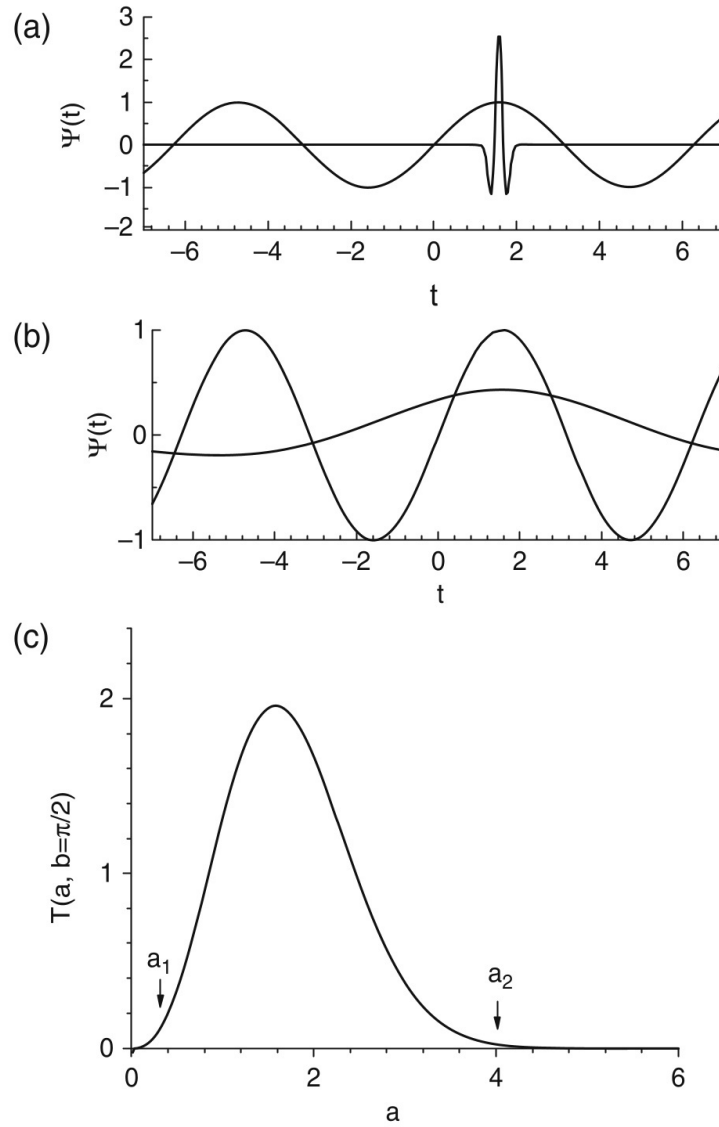
Having defined the wavelet and its transform, we are ready to see how the transform is used as a signal analysis tool. In plain words, the wavelet transform works as a mathematical microscope, where  $b$  is the location on the time series being viewed and  $a$  represents the magnification at location  $b$ .

Let us look at a simple example evaluating the wavelet transform  $T(a, b)$ . Figures 14.3 and 14.4 show the same sinusoidal waves together with Mexican hat wavelets of various locations and dilations. In Fig. 14.3a, the wavelet is located on a segment of the signal on which a positive part of the signal is fairly coincidental with that of the wavelet. This results in a large positive



**Fig. 14.3.** (a), (b) Positional relations between the wavelet (*thick*) and signal (*thin*). The wavelet in (a) located at  $b_1 = \pi/2$  is in phase with the signal, which results in a large positive value of  $T(a, b)$  at  $b_1$ . The wavelet in (b) located at  $b_2 = -\pi/2$  is out of phase with the signal, which yields a large negative value of  $T(b)$  at  $b_2$ . (c) The plot of  $T(a = 1.0, b)$  as a function of  $b$

value of  $T(a, b)$  in (14.6). In Fig. 14.3b, the wavelet is moved to a new location where the wavelet and the signal are out of phase. In this case, the convolution expressed by (14.6) produces a large negative value of  $T(a, b)$ . In between these two extrema, the value of  $T(a, b)$  decreases from a maximum to a minimum as shown in Fig. 14.3. The three figures thus clearly demonstrate how the wavelet transform  $T(a, b)$  depends on the translation parameter  $b$  of the wavelet of interest.



**Fig. 14.4.** Wavelets with  $a = 0.33$  (a) and  $a = 4.0$  (b), in which  $b = \pi/2$  is fixed. The resulting wavelet transform  $T(a, b = \pi/2)$  as a function of  $a$  is given in (c)

In a similar way, Fig. 14.4 a–c shows the dependence of  $T(a, b)$  on the dilatation parameter  $a$ . When  $a$  is quite small, the positive and negative parts of the wavelet are all convolved by roughly the same part of the signal  $x(t)$ , producing a value of  $T(a, b)$  near zero (see Fig. 14.4a). Likewise,  $T(a, b)$  tends to zero as  $a$  becomes very large (see Fig. 14.4b), since the wavelet covers many positive and negatively repeating parts of the signal. These latter two results indicate that when the dilatation parameter  $a$  is either very small or very large compared with the period of the signal, the wavelet transform  $T(a, b)$  gives near-zero values.

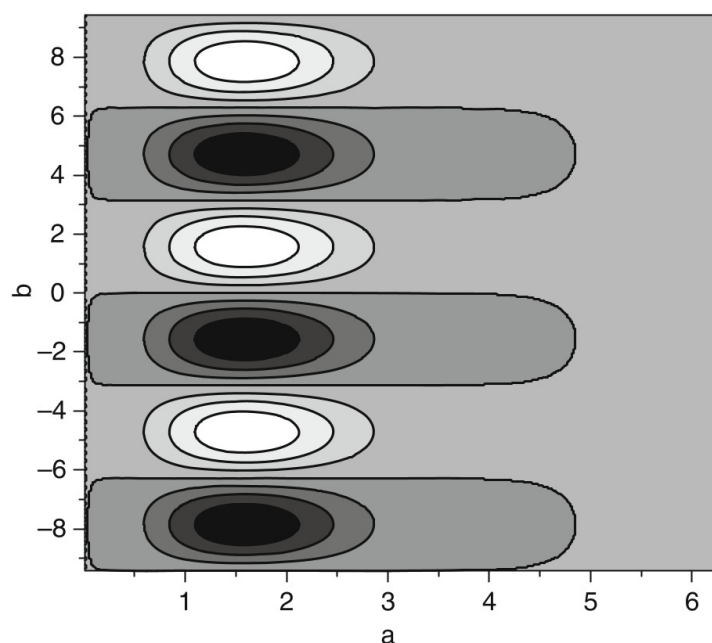
Figure 14.5 shows a contour plot of  $T(a, b)$  vs.  $a$  and  $b$  for a sinusoidal signal

$$x(t) = \sin t,$$

where the Mexican hat wavelet has been used. The light and shadowed regions indicate positive and negative magnitudes of  $T(a, b)$ , respectively. The near-zero values of  $T(a, b)$  are evident in the plot at both large and small values of  $a$ . In addition, at intermediate values of  $a$ , we observe large undulations in  $T(a, b)$  corresponding to the sinusoidal form of the signal. This wavelike behavior is accounted for by referring back to Figs. 14.3a–b and 14.4a–b, where wavelets move in and out of phase with the signal.

Therefore, when the wavelet matches the shape of the signal well at a specific scale and location, the transform value is high. On the other hand, if the wavelet and the signal do not correlated well, the transform value is low. Carrying out the process at various signal locations and for various wavelet scales, we can determine the correlation between the wavelet and the signal.

**Remark.** In Fig. 14.5, the maxima and minima of the transform occur at an  $a$  scale of one quarter of the period,  $\pi/2$ , of the sine wave  $x(t) = \sin t$ . This feature holds in general; correlation between the wavelet  $\psi_{a,b}(t)$  and the signal  $x(t)$  with a period  $p$  becomes a maximum at  $a = p/4$ .

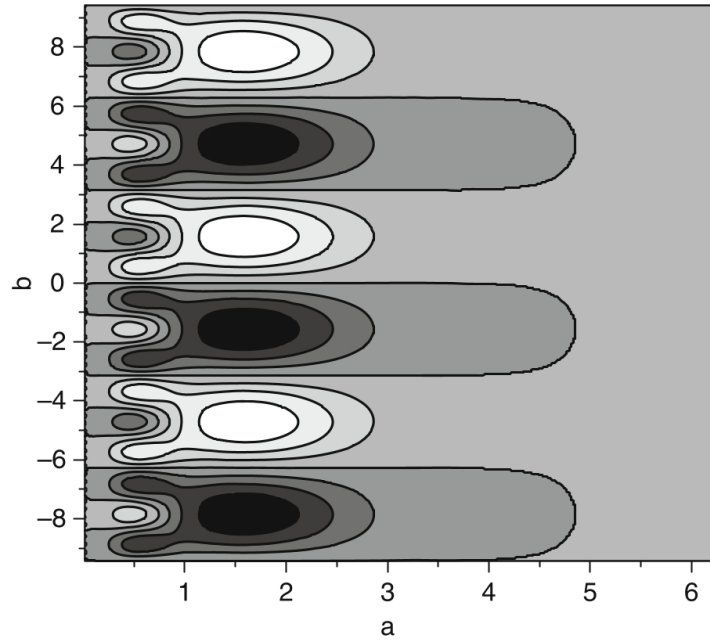


**Fig. 14.5.** Contour plot of the wavelet transform  $T(a, b)$  of a sinusoidal wave  $x(t) = \sin t$

#### 14.1.4 Actual Application of the Wavelet Transform

The wavelet transformation procedure can be applied to signals that have a more complicated wave form than a simple sinusoidal wave. Figure 14.6 shows a signal

$$x(t) = \sin t + \sin 3t$$



**Fig. 14.6.** Wavelet transform  $T(a, b)$  of a complicated signal  $x(t) = \sin t + \sin 3t$

composed of two sinusoidal waves with different frequencies. The wavelet transform  $T(a, b)$  of  $x(t)$  is plotted in Fig. 14.6. It is clear that the contribution from the wave with the higher-frequency oscillation appears at a smaller  $a$  scale. This clearly demonstrates the ability of the wavelet transform to decompose the original signal into its separate components.

#### 14.1.5 Inverse Wavelet Transform

Similar to its Fourier counterpart, there is an **inverse wavelet transformation**, that enables us to reproduce the original signal  $x(t)$  from its wavelet transform  $T(a, b)$ .

♠ **Inverse wavelet transform:**

If  $x \in L^2(\mathbf{R})$ , then  $f$  can be reconstructed by equation

$$x(t) = \frac{1}{C_\Psi} \int_{-\infty}^{\infty} db \int_0^{\infty} \frac{da}{a^2} T(a, b) \psi_{a,b}(t), \quad (14.7)$$

where the equality holds **almost everywhere**.

The proof of the equation is based on the lemma below.

♠ **Parseval identity for wavelet transform:**

Let  $T_f(a, b), T_g(a, b)$  be the wavelet transform of  $f(t), g(t) \in L^2(\mathbf{R})$ , respectively, associated with the wavelet  $\psi_{a,b}(t)$ . Then we have

$$\int_0^\infty \frac{da}{a^2} \int_{-\infty}^\infty db T_f(a, b) T_g^*(a, b) = C_\Psi \int_{-\infty}^\infty f(t) g(t)^* dt. \quad (14.8)$$

This identity is derived in Exercise 4. We are now ready to prove the inverse transformation (14.7).

**Proof (of the inverse wavelet transformation):** Assume an arbitrary real function  $g(t) \in L^2(\mathbf{R})$ . It follows from the Parseval identity that

$$\begin{aligned} C_\Psi \int_{-\infty}^\infty f(t) g(t) dt &= \int_{-\infty}^\infty db \int_0^\infty \frac{da}{a^2} T_f(a, b) T_g(a, b) \\ &= \int_{-\infty}^\infty db \int_0^\infty \frac{da}{a^2} T_f(a, b) \int_{-\infty}^\infty g(t) \psi_{a,b}(t) dt \\ &= \int_{-\infty}^\infty dt g(t) \left[ \int_{-\infty}^\infty db \int_0^\infty \frac{da}{a^2} T_f(a, b) \psi_{a,b}(t) \right]. \end{aligned}$$

Since  $g(t)$  is arbitrary, the inverse equation (14.7) follows. ♣

#### 14.1.6 Noise Reduction Technique

Suppose that the inverse transformation equation (14.7) is rewritten as

$$x^*(t) = \frac{1}{C_\Psi} \int_{-\infty}^\infty db \int_{a^*}^\infty \frac{da}{a^2} T(a, b) \psi_{a,b}(t),$$

the integration range with respect to  $a$  in an interval  $[a^*, \infty)$  with  $a^* > 0$ . Then, the result  $x^*(t)$  obtained on the left-hand side deviates from the original signal  $x(t)$  owing to the lack of information for the scale from  $a = 0$  to  $a = a^*$ . In applications, this deviation property is made use of as a noise reduction technique.

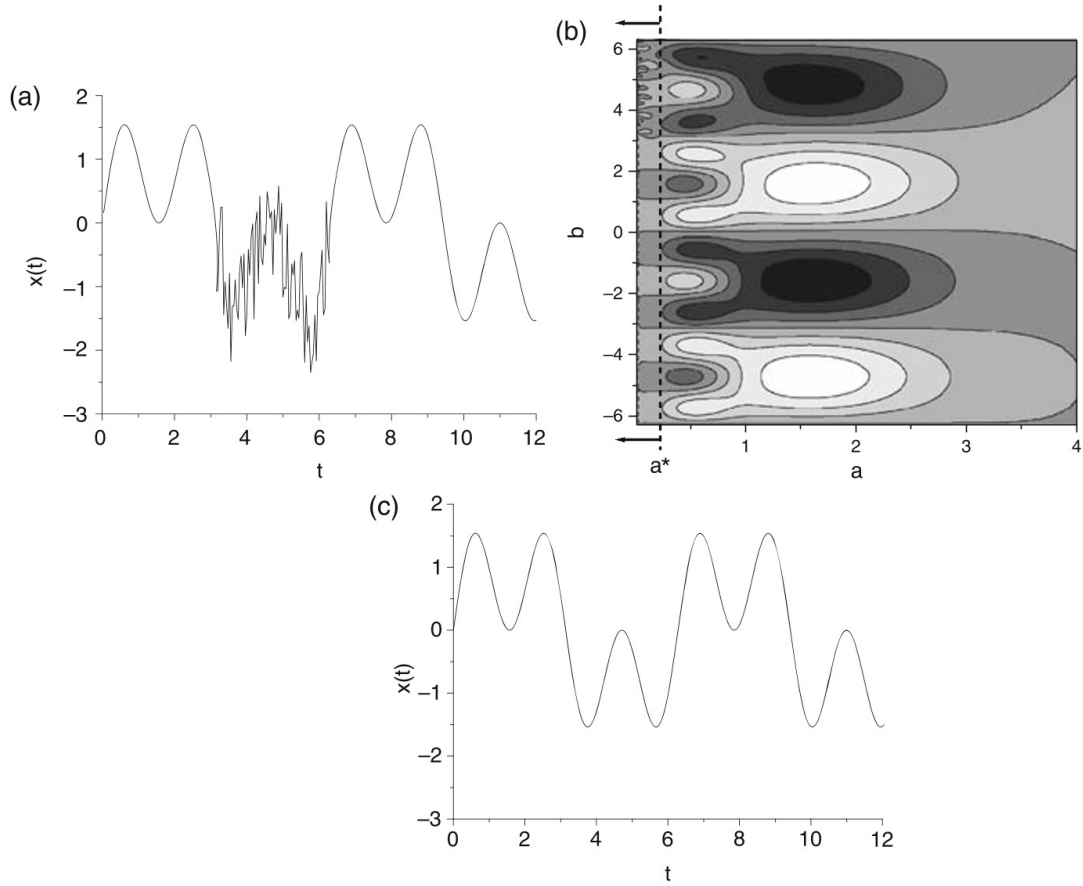
By way of a demonstration, Fig. 14.7a illustrates a segment of the signal

$$x(t) = \sin t + \sin 3t + R(t)$$

constructed from two sinusoidal waveforms plus a local burst of noise  $R(t)$ . The transform plot of the composite signal shows the two constituent waveforms at scales  $a_1 = \pi/2$  and  $a_2 = \pi/6$  in addition to a burst of noise around  $b = 5.0$  in a high-frequency region (i.e., small  $a$  scale).

Now we try to remove the high-frequency noise component by means of the following reconstruction procedure. Figure 14.7b shows a reconstruction of the signal where we artificially set  $T(a, b) = 0$  for  $a < a^*$ . In effect, we are reconstructing the signal using

$$x(t) = \frac{1}{C_\Psi} \int_{-\infty}^\infty db \int_{a^*}^\infty \frac{da}{a^2} T(a, b) \psi_{a,b}(t),$$



**Fig. 14.7.** Noise reduction procedure through wavelet transformation. **(a)** A signal  $x(t) = \sin t + \sin 3t + R(t)$  with a local burst of noise  $R(t)$ . **(b)** The wavelet transform  $T(a, b)$  of the  $x(t)$ . Noise reduction is accomplished through the inverse transformation of the  $T(a, b)$  by applying an artificial condition of  $T(a < a^*, b) = 0$ . **(c)** The reconstructed signal  $x^*(t)$  from the noise-reduction procedure

i.e., over a range of scales  $[a^*, \infty)$ . The lower integral limit,  $a^*$ , is the cut-off scale indicated by the dotted line in Fig. 14.7b. As a result, the high-frequency noise component evidently reduces in the reconstructed signal as shown in Fig. 14.7c. This simple noise reduction method is known as **scale-dependent thresholding**.

## Exercises

1. Show that the Fourier transform of the Haar wavelet satisfies the admissible condition (14.1).

**Solution:** The Fourier transform  $\Psi(\omega)$  of the Haar wavelet  $\psi(t)$  is given by

$$\Psi(\omega) = \int_0^{1/2} e^{-i\omega t} dt - \int_{1/2}^1 e^{-i\omega t} dt = ie^{-i\omega/2} \frac{\sin^2(\omega/4)}{\omega/4}.$$

Hence, we have

$$C_\Psi = \int_0^\infty \frac{|\Psi(\omega)|^2}{\omega} d\omega = 16 \int_0^\infty \frac{\sin^4(\omega/4)}{\omega^3} d\omega < \infty. \quad \clubsuit$$

2. Prove that the Fourier transform of  $\psi_{a,b}(t)$  yields  $\Psi_{a,b}(\omega) = \sqrt{a}e^{-ib\omega}\Psi(a\omega)$ .

**Solution:** It readily follows that

$$\Psi_{a,b}(\omega) = \frac{1}{\sqrt{a}} \int_{-\infty}^\infty e^{-i\omega t} \psi_{a,b}(t) dt = \frac{1}{\sqrt{a}} \int_{-\infty}^\infty e^{-i\omega t} \psi\left(\frac{t-b}{a}\right) dt.$$

Set  $u = (t-b)/a$  in the last integral to obtain

$$\Psi_{a,b}(\omega) = \frac{1}{\sqrt{a}} \int_{-\infty}^\infty e^{-i\omega(au+b)} \psi(u) a du = \sqrt{a}e^{-ib\omega}\Psi(a\omega). \quad \clubsuit$$

3. Let  $\psi(t)$  be a wavelet and  $\phi(t)$  be a real, bounded, and integrable function. Show that the convolution  $\psi * \phi$  is also a wavelet.

**Solution:** We first show that  $\psi * \phi \in L^2(\mathbf{R})$ . Observe that

$$\begin{aligned} [\psi(t) * \phi(t)]^2 &= \left[ \int_{-\infty}^\infty \psi(t-u)\phi(u) du \right]^2 \\ &= \left[ \int_{-\infty}^\infty \psi(t-u) \phi(u)^{1/2} \phi(u)^{1/2} du \right]^2 \\ &\leq \int_{-\infty}^\infty \psi(t-u)^2 \phi(u) du \int_{-\infty}^\infty \phi(u') du'. \end{aligned}$$

The integral  $\int_{-\infty}^\infty \phi(u') du'$  is a constant, denoted by  $A$ . Integrate both sides with respect to  $t$  to obtain

$$\begin{aligned} \int_{-\infty}^\infty [\psi(t) * \phi(t)]^2 dt &\leq A \int_{-\infty}^\infty \phi(u) \left[ \int_{-\infty}^\infty \psi(t-u)^2 dt \right] du \\ &= A \int_{-\infty}^\infty \phi(u) du \int_{-\infty}^\infty \psi(t)^2 dt = A^2 \int_{-\infty}^\infty \psi(t)^2 dt < \infty, \end{aligned}$$

which clearly indicates that  $\psi * \phi \in L^2(\mathbf{R})$ . Next we show that the convolution  $\psi * \phi$  satisfies the admissibility condition. In fact,

$$\begin{aligned} \int_{-\infty}^\infty \frac{|\mathcal{F}[\psi * \phi]|^2}{\omega} d\omega &= \int_{-\infty}^\infty \frac{|\Psi(\omega)\Phi(\omega)|^2}{\omega} d\omega \\ &= \int_{-\infty}^\infty \frac{|\Psi(\omega)|^2}{\omega} \sup |\Phi(\omega)|^2 d\omega < \infty. \end{aligned}$$

These two results imply that the convolution  $\psi * \phi$  is a wavelet.  $\clubsuit$

4. Derive the Parseval identity for the wavelet transform (14.8).

**Solution:** The transform  $T_f(a, b)$  reads

$$T_f(a, b) = \int_{-\infty}^{\infty} f(t) \psi_{a,b}^*(t) dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) \sqrt{a} e^{-ib\omega} \Psi(a\omega) d\omega,$$

where we used the fact that  $\Psi_{a,b}(\omega) = \sqrt{a} e^{-ib\omega} \Psi(a\omega)$ . Similarly, we have  $T_g(a, b) = \frac{1}{2\pi} \int_{-\infty}^{\infty} G(\omega) \sqrt{a} e^{-ib\omega} \Psi(a\omega) d\omega$ . Hence, we have

$$\begin{aligned} & \int_0^{\infty} \frac{da}{a^2} \int_{-\infty}^{\infty} db T_f(a, b) T_g(a, b) \\ &= \int_0^{\infty} \frac{da}{a^2} \int_{-\infty}^{\infty} db \int_{-\infty}^{\infty} d\omega \int_{-\infty}^{\infty} d\omega' \frac{a e^{-ib(\omega+\omega')}}{(2\pi)^2} F(\omega) G(\omega') \Psi(a\omega) \Psi(a\omega') \\ &= \frac{1}{2\pi} \int_0^{\infty} \frac{da}{a} \int_{-\infty}^{\infty} d\omega \int_{-\infty}^{\infty} d\omega' F(\omega) G(\omega') \Psi(a\omega) \Psi(a\omega') \delta(\omega + \omega') \\ &= \frac{1}{2\pi} \int_0^{\infty} \frac{da}{a} \int_{-\infty}^{\infty} d\omega F(\omega) G(-\omega) \Psi(a\omega) \Psi(-a\omega). \end{aligned}$$

Since  $\psi(t)$  and  $g(t)$  are both real,  $\Psi(-a\omega) = \Psi(a\omega)^*$  and  $G(-\omega) = G^*(\omega)$ . Thus we have

$$\begin{aligned} \int_0^{\infty} \frac{da}{a^2} \int_{-\infty}^{\infty} db T_f(a, b) T_g(a, b) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} d\omega F(\omega) G^*(\omega) \int_0^{\infty} \frac{|\Psi(x)|^2}{x} dx \\ &= C_{\Psi} \int_{-\infty}^{\infty} f(t) g(t) dt, \end{aligned}$$

where  $x \equiv a\omega$ . This completes the proof. ♣

## 14.2 Discrete Wavelet Analysis

### 14.2.1 Discrete Wavelet Transforms

Having discussed the continuous wavelet transform, we move on to its discrete version, known as the **discrete wavelet transform**. In many applications, data are represented by a finite number of values, so it is important and often useful to consider the discrete version of a wavelet transform. We also can use an efficient numerical algorithm, called the **fast wavelet transform**, which allows us to compute the wavelet transform of the signal and its inverse quite efficiently.

We begin with the definition of a **discrete wavelet**. In the previous section, the wavelet function was defined at scale  $a$  and location  $b$  as

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{a}\right),$$

in which the values of parameters  $a$  and  $b$  can change continuously. We now want to discretize the values of  $a$  and  $b$ . One possible way to sample  $a$  and  $b$  is to use a logarithmic discretization of the  $a$  scale and link this to the size of the steps taken between  $b$  locations. This kind of discretization yields

$$\psi_{m,n}(t) = \frac{1}{\sqrt{a_0^m}} \psi\left(\frac{t - nb_0 a_0^m}{a_0^m}\right), \quad (14.9)$$

where the integers  $m$  and  $n$  control the wavelet dilation and translation respectively;  $a_0$  is a specified fixed dilation step parameter and  $b_0$  is the location parameter. In the expression (14.9), the size of the translation steps,  $\Delta b = b_0 a_0^m$ , is directly proportional to the wavelet scale,  $a_0^m$ .

Common choices for discrete wavelet parameters  $a_0$  and  $b_0$  are  $1/2$  and  $1$ , respectively. This power-of-two logarithmic scaling of the dilation steps is known as the **dyadic grid arrangement**. Substituting  $a_0 = 1/2$  and  $b_0 = 1$  into (14.9), we obtain the **dyadic grid wavelet** represented by

$$\psi_{m,n}(t) = 2^{m/2} \psi(2^m t - n). \quad (14.10)$$

Using the dyadic grid wavelet of (14.10), we arrive at the **discrete wavelet transform** of a continuous signal  $x(t)$ :

♠ **Discrete wavelet transform:**

$$T_{m,n} = \int_{-\infty}^{\infty} x(t) \psi_{m,n}(t) dt = \int_{-\infty}^{\infty} x(t) 2^{m/2} \psi(2^m t - n) dt. \quad (14.11)$$

*Remark.* Note that the discrete wavelet transform (14.11) differs from the discretized approximation of the continuous wavelet transform given by

$$T(a,b) = \int_{-\infty}^{\infty} x(t) \psi_{a,b}^*(t) dt \simeq \sum_{l=-\infty}^{\infty} x(l\Delta t) \psi_{a,b}^*(l\Delta t) \Delta t. \quad (14.12)$$

In (14.12), the integration variable  $t$  is discretized, and  $a$  and  $b$  are continuous whose values can be arbitrarily chosen. On the other hand, in the discrete wavelet transform (14.11),  $a$  and  $b$  are discretized and  $t$  remains continuous.

### 14.2.2 Complete Orthonormal Wavelets

The fundamental question is whether the original signal  $x(t)$  can be constructed from the discrete wavelet transform  $T_{m,n}$  through the relation

$$x(t) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} T_{m,n} \psi_{m,n}(t). \quad (14.13)$$

As intuitively understood, the reconstruction equation (14.13) is justified if the discretized wavelets  $\psi_{m,n}(t)$  are **orthonormal** and **complete**. The completeness of  $\psi_{m,n}(t)$  implies that any function  $x \in L^2(\mathbf{R})$  can be expanded by

$$x(t) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} c_{m,n} \psi_{m,n}(t) \quad (14.14)$$

with appropriate expansion coefficients  $c_{m,n}$ . Hence, the orthonormality

$$\int_{-\infty}^{\infty} \psi_{m,n}(t) \psi_{m',n'}(t) dt = \delta_{m,n} \delta_{m',n'} \quad (14.15)$$

results in  $c_{m,n} = T_{m,n}$  in (14.14) because

$$\begin{aligned} T_{m,n} &= \int_{-\infty}^{\infty} x(t) \psi_{m,n}(t) dt = \int_{-\infty}^{\infty} \left[ \sum_{m'=-\infty}^{\infty} \sum_{n'=-\infty}^{\infty} c_{m',n'} \psi_{m',n'}(t) \right] \psi_{m,n}(t) dt \\ &= \sum_{m'=-\infty}^{\infty} \sum_{n'=-\infty}^{\infty} c_{m',n'} \int_{-\infty}^{\infty} \psi_{m,n}(t) \psi_{m',n'}(t) dt \\ &= \sum_{m'=-\infty}^{\infty} \sum_{n'=-\infty}^{\infty} c_{m',n'} \delta_{m,n} \delta_{m',n'} = c_{m,n}. \end{aligned}$$

In general, however, the wavelets  $\psi_{m,n}(t)$  given by (14.9) are neither orthonormal nor complete. We thus arrive at the following theorem:

♠ **Validity of the inverse transformation formula:**

The inverse transformation formula (14.13) is valid only for a limited class of sets of discrete wavelets  $\{\psi_{m,n}(t)\}$  that is endowed with both orthonormality and completeness.

The simplest example of such desired wavelets is the **Haar discrete wavelet** presented below.

*Examples* The Haar discrete wavelet is defined by

$$\psi_{m,n}(t) = 2^{m/2} \psi(2^m t - n),$$

where

$$\psi(t) = \begin{cases} 1 & 0 \leq t < 1/2, \\ -1 & 1/2 \leq t < 1, \\ 0 & \text{otherwise.} \end{cases}$$

This wavelet is known to be orthonormal and complete; its orthonormality is verified in Exercise 1.

### 14.2.3 Multiresolution Analysis

We know from Sect. 14.2.2 that in order to use equation (14.13), we must find an appropriate set of discrete wavelets  $\{\psi_{m,n}\}$  that possess both orthonormality and completeness. In the remainder of this section, we describe a framework for constructing such discrete wavelets that is based on the concept of **multiresolution analysis**.

Multiresolution analysis involves a particular class of a set of function spaces. The greatest peculiarity is that it establishes a nesting structure of subspaces of  $L^2(\mathbf{R})$  that allows us to construct a complete orthonormal set of functions (i.e., an **orthonormal basis**) for  $L^2(\mathbf{R})$ . The resulting orthonormal basis is simply the discrete wavelet  $\psi_{m,n}(t)$  that yields the reconstruction equation (14.13).

♠ **Multiresolution analysis:** A multiresolution analysis involves a set of function spaces that consists of a sequence  $\{\mathcal{V}_j : j \in \mathbf{Z}\}$  of closed subspaces of  $L^2(\mathbf{R})$ . Here the subspaces  $\mathcal{V}_j$  satisfy the following conditions:

1.  $\cdots \subset \mathcal{V}_{-2} \subset \mathcal{V}_{-1} \subset \mathcal{V}_0 \subset \mathcal{V}_1 \subset \mathcal{V}_2 \cdots \subset L^2(\mathbf{R})$ .
2.  $\bigcap_{j=-\infty}^{\infty} \mathcal{V}_j = \{0\}$ .
3.  $f(t) \in \mathcal{V}_j$  if and only if  $f(2t) \in \mathcal{V}_{j+1}$  for all integers  $j$ .
4. There exists a function  $\phi(t) \in \mathcal{V}_0$  such that the set  $\{\phi(t - n), n \in \mathbf{Z}\}$  is an orthonormal basis for  $\mathcal{V}_0$ .

The function  $\phi(t)$  introduced above is called the **scaling function** (or **father wavelet**). It should be emphasized that the above definition gives no information as to the existence of (or the way to construct) the function  $\phi(t)$  satisfying condition 4. However, once we find such a function  $\phi(t)$ , we can establish a multiresolution analysis  $\{\mathcal{V}_j\}$  by defining the function space  $\mathcal{V}_0$  spanned by the orthonormal basis  $\{\phi(t - n), n \in \mathbf{Z}\}$  and then forming other subspaces  $\mathcal{V}_j$  ( $j \neq 0$ ) successively by using the property denoted in condition 3. If this is achieved, we say that *our scaling function  $\phi(t)$  generates the multiresolution analysis  $\{\mathcal{V}_j\}$* .

**Remark.** There is no straightforward way to construct a scaling function  $\phi(t)$  or, equivalently, a multiresolution analysis  $\{\mathcal{V}_j\}$ . Nevertheless, many kinds of scaling functions have been discovered by means of sophisticated mathematical techniques. Here we omit the details of the derivations and just refer to the resulting scaling function at need.

**Examples** Consider the space  $\mathcal{V}_m$  of all functions in  $L^2(\mathbf{R})$  that are constant in each interval  $[2^{-m}n, 2^{-m}(n+1)]$  for all  $n \in \mathbf{Z}$ . Obviously, the space  $\mathcal{V}_m$  satisfies conditions 1–3 of a multiresolution analysis. Furthermore, it is easy to see that the set  $\{\phi(t-n), n \in \mathbf{Z}\}$  depicted in Fig. 14.8, which is defined by

$$\phi(t) = \begin{cases} 1, & 0 \leq t \leq 1, \\ 0, & \text{otherwise,} \end{cases} \quad (14.16)$$

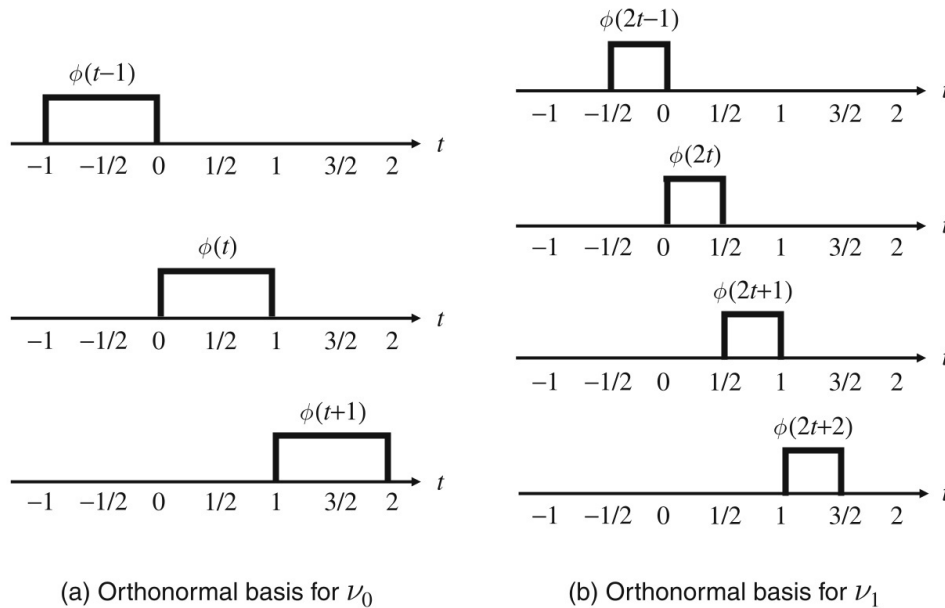
satisfies condition 4. Hence, any function  $f \in \mathcal{V}_0$  can be expressed by

$$f(t) = \sum_{n=-\infty}^{\infty} c_n \phi(t-n),$$

with appropriate constants  $c_n$ . Thus, the spaces  $\mathcal{V}_m$  consist of the multiresolution analysis generated by the scaling function (14.16).

#### 14.2.4 Orthogonal Decomposition

The importance of a multiresolution analysis lies in its ability to construct an orthonormal basis (i.e., a complete orthonormal set of functions) for  $L^2(\mathbf{R})$ .



**Fig. 14.8.** Two different sets of functions:  $\mathcal{V}_0$  and  $\mathcal{V}_1$

In order to prove this statement, we first recall that a multiresolution analysis  $\{\mathcal{V}_j\}$  satisfies the relation

$$\mathcal{V}_0 \subset \mathcal{V}_1 \subset \mathcal{V}_2 \subset \cdots \subset L^2.$$

We now define a space  $\mathcal{W}_0$  as the **orthogonal complement** of  $\mathcal{V}_0$  and  $\mathcal{V}_1$ , which yields

$$\mathcal{V}_1 = \mathcal{V}_0 \oplus \mathcal{W}_0. \quad (14.17)$$

The space  $\mathcal{W}_0$  we have introduced is called the **wavelet space** of zero order: the reason for the name is clarified in Sect. 14.2.5. The relation (14.17) extends to

$$\mathcal{V}_2 = \mathcal{V}_1 \oplus \mathcal{W}_1 = \mathcal{V}_0 \oplus \mathcal{W}_0 \oplus \mathcal{W}_1 \quad (14.18)$$

or, more generally, it gives

$$L^2 = \mathcal{V}_\infty = \mathcal{V}_0 \oplus \mathcal{W}_0 \oplus \mathcal{W}_1 \oplus \mathcal{W}_2 \oplus \cdots, \quad (14.19)$$

where  $\mathcal{V}_0$  is the initial space spanned by the set of functions  $\{\phi(t-n), n \in \mathbb{Z}\}$ . Figure 14.9 illustrates the nesting structure of the spaces  $\mathcal{V}_j$  and  $\mathcal{W}_j$  for different scales  $j$ .

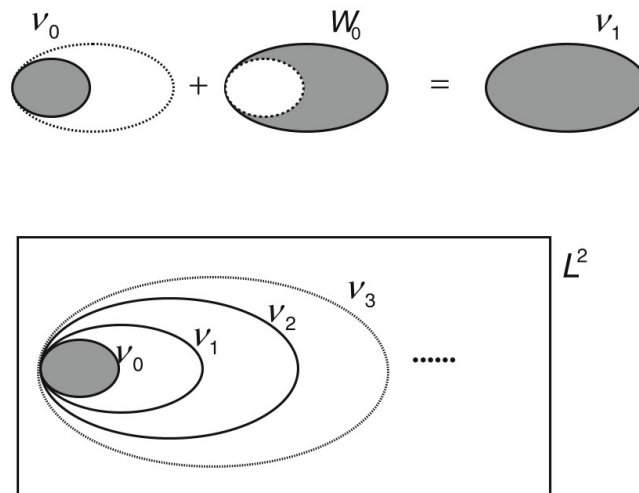
Since the scale of the initial space is arbitrary, it can be chosen at a higher resolution such as

$$L^2 = \mathcal{V}_5 \oplus \mathcal{W}_5 \oplus \mathcal{W}_6 \oplus \cdots,$$

or at a lower resolution such as

$$L^2 = \mathcal{V}_{-3} \oplus \mathcal{W}_{-3} \oplus \mathcal{W}_{-2} \oplus \cdots,$$

or even at negative infinity, where (14.19) becomes



**Fig. 14.9.** Hierarchical structure of the spaces  $\mathcal{V}_j$  and  $\mathcal{W}_j$  as subspaces of  $L^2$

$$L^2 = \cdots \oplus \mathcal{W}_{-1} \oplus \mathcal{W}_0 \oplus \mathcal{W}_1 \oplus \cdots . \quad (14.20)$$

The expression (14.20) is referred to as the **orthogonal decomposition** of the  $L^2$  space and indicates that any function  $x \in L^2(\mathbf{R})$  can be decomposed into the infinite sum of  $g_j \in \mathcal{W}_j$ :

$$x(t) = \cdots + g_{-1}(t) + g_0(t) + g_1(t) + \cdots . \quad (14.21)$$

#### 14.2.5 Constructing an Orthonormal Basis

Let us further examine the orthogonal property of the wavelet spaces  $\{\mathcal{W}_j\}$ . From (14.17) and (14.18), we have

$$\mathcal{W}_0 \subset \mathcal{V}_1 \text{ and } \mathcal{W}_1 \subset \mathcal{V}_2.$$

In view of the definition of the multiresolution analysis  $\{\mathcal{V}_j\}$ , it follows that

$$f(t) \in \mathcal{V}_1 \iff f(2t) \in \mathcal{V}_2,$$

so

$$f(t) \in \mathcal{W}_0 \iff f(2t) \in \mathcal{W}_1. \quad (14.22)$$

Furthermore, condition 4 in Sect. 14.2.3 results in

$$f(t) \in \mathcal{W}_0 \iff f(t - n) \in \mathcal{W}_0 \text{ for any } n \in \mathbf{Z}. \quad (14.23)$$

The two results (14.22) and (14.23) are ingredients for constructing the orthonormal basis of  $L^2(R)$  that we are looking for, as demonstrated below.

We first assume that there exists a function  $\psi(t)$  that leads to an orthonormal basis  $\{\psi(t - n), n \in \mathbf{Z}\}$  for the space  $\mathcal{W}_0$ . Then, if we use the notation

$$\psi_{0,n}(t) \equiv \psi(t - n) \in \mathcal{W}_0,$$

it follows from (14.22) and (14.23) that its scaled version defined by

$$\psi_{1,n}(t) = \sqrt{2}\psi(2t - n)$$

serves as an orthonormal basis for  $\mathcal{W}_1$ . The term  $\sqrt{2}$  was introduced to keep the normalization condition

$$\int_{-\infty}^{\infty} \psi_{0,n}(t)^2 dt = \int_{-\infty}^{\infty} \psi_{1,n}(t)^2 dt = 1.$$

By repeating the same procedure, we find that the function

$$\psi_{m,n}(t) = 2^{m/2}\psi(2^m t - n) \quad (14.24)$$

constitutes an orthonormal basis for the space  $\mathcal{W}_m$ . Applying these results to the expression (14.21), we have for any  $x \in L^2(\mathbf{R})$ ,

$$\begin{aligned} x(t) &= \cdots + g_{-1}(t) + g_0(t) + g_1(t) + \cdots \\ &= \cdots + \sum_{n=-\infty}^{\infty} c_{-1,n} \psi_{-1,n}(t) + \sum_{n=-\infty}^{\infty} c_{0,n} \psi_{0,n}(t) + \sum_{n=-\infty}^{\infty} c_{1,n} \psi_{1,n}(t) + \cdots \\ &= \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} c_{m,n} \psi_{m,n}(t). \end{aligned} \quad (14.25)$$

Hence, the family  $\psi_{m,n}(t)$  represents an orthonormal basis for  $L^2(\mathbf{R})$ . The above arguments are summarized by the following theorem:

♠ **Theorem:**

Let  $\{\mathcal{V}_j\}$  be a multiresolution analysis and define the space  $\mathcal{W}_0$  by  $\mathcal{W}_0 = \mathcal{V}_1 \setminus \mathcal{V}_0$ . If a function  $\psi(t)$  that leads to an orthonormal basis  $\{\psi(t-n), n \in \mathbf{Z}\}$  for  $\mathcal{W}_0$  is found, then the set of functions  $\{\psi_{m,n}, m, n \in \mathbf{Z}\}$  given by

$$\psi_{m,n}(t) = 2^{m/2} \psi(2^m t - n)$$

constitutes an orthonormal basis for  $L^2(\mathbf{R})$ .

Emphasis is placed on the fact that since  $\psi_{m,n}(t)$  is the orthonormal basis for  $L^2(\mathbf{R})$ , the coefficients  $c_{m,n}$  in (14.25) are identical to the discrete wavelet transform  $T_{m,n}$  given by (14.11) (see Sect. 14.2.2). Therefore, the function  $\psi(t)$  we introduce here is identified with the wavelet in the framework of continuous and discrete wavelet analysis, such as the Haar and the Mexican hat wavelets. In this sense, each  $\mathcal{W}_m$  is referred to as the **wavelet space** and the function  $\psi(t)$  is sometimes called the **mother wavelet**.

### 14.2.6 Two-Scale Relations

The preceding argument suggests that an orthonormal basis  $\{\psi_{m,n}\}$  for  $L^2(\mathbf{R})$  can be constructed by specifying the explicit function form of the mother wavelet  $\psi(t)$ . Thus the remaining task is to develop a systematic way of determining the mother wavelet  $\psi(t)$  that leads to an orthonormal basis  $\{\psi(t-n) \mid n \in \mathbf{Z}\}$  for the space  $\mathcal{W}_0 = \mathcal{V}_1 \setminus \mathcal{V}_0$  contained in a given multiresolution analysis. We shall see that the  $\psi(t)$  can be found by examining the properties of the scaling function  $\phi(t)$ ; we should recall that  $\phi(t)$  yields an orthonormal basis  $\{\phi(t-n) \mid n \in \mathbf{Z}\}$  for the space  $\mathcal{V}_0$ . (In this context, the space  $\mathcal{V}_j$  is sometimes referred to as the **scaling function space**.)

In this subsection, we make reference to an important feature of the scaling function  $\phi(t)$  called the **two-scale relation**, which plays a key role in constructing the mother wavelet  $\psi(t)$  of a given multiresolution analysis. We already know that all the functions in  $\mathcal{V}_m$  are obtained from those in  $\mathcal{V}_0$  through scaling by  $2^m$ . Applying this result to the scaling function denoted by

$$\phi_{0,n}(t) \equiv \phi(t - n) \in \mathcal{V}_0,$$

we find that

$$\phi_{m,n}(t) = 2^{m/2} \phi(2^m t - n), \quad m \in \mathbf{Z} \quad (14.26)$$

is an orthonormal basis for  $\mathcal{V}_m$ . In particular, since  $\phi \in \mathcal{V}_0 \subset \mathcal{V}_1$  and  $\phi_{1,n}(t) = \sqrt{2} \phi(2t - n)$  is an orthonormal basis for  $\mathcal{V}_1$ ,  $\phi(t)$  can be expanded by  $\phi_{1,n}(t)$ . This is formally stated in the following theorem:

♠ **Two-scale relation:**

If the scaling function  $\phi(t)$  generates a multiresolution analysis  $\{\mathcal{V}_j\}$ , it satisfies the recurrence relation:

$$\phi(t) = \sum_{n=-\infty}^{\infty} p_n \phi_{1,n}(t) = \sqrt{2} \sum_{n=-\infty}^{\infty} p_n \phi(2t - n), \quad (14.27)$$

where

$$p_n = \int_{-\infty}^{\infty} \phi(t) \phi_{1,n}(t) dt. \quad (14.28)$$

This recurrence equation is called the **two-scale relation** of  $\phi(t)$  and the coefficients  $p_n$  are called the **scaling function coefficients**.

**Remark.** The two-scale relation is also referred to as the **multiresolution analysis equation**, the **refinement equation**, or the **dilation equation**, depending on the context.

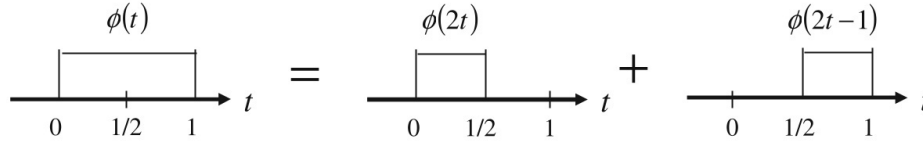
*Examples* Consider again the space  $\mathcal{V}_m$  of all functions in  $L^2(\mathbf{R})$  that are constant on intervals  $[2^{-m}n, 2^{-m}(n+1)]$  with  $n \in \mathbf{Z}$ . This multiresolution analysis is known to be generated by the scaling function  $\phi(t)$  of (14.16). Substituting (14.16) into (14.28), we obtain

$$p_0 = p_1 = \frac{1}{\sqrt{2}} \quad \text{and} \quad p_n = 0 \quad \text{for } n \neq 0, 1.$$

Thus the two-scale relation reads

$$\phi(t) = \phi(2t) + \phi(2t - 1).$$

This means that the scaling function  $\phi(t)$  in this case is a linear combination of its contracted versions as depicted in Fig. 14.10.



**Fig. 14.10.** Two-scale relation of  $\phi(t)$

### 14.2.7 Constructing the Mother Wavelet

We are now in a position to determine the mother wavelet  $\psi(t)$  that enables us to establish an orthonormal basis  $\{\psi(t-n), n \in \mathbf{Z}\}$  for  $L^2(\mathbf{R})$ . Recall that a mother wavelet  $\psi(t) = \psi_{0,0}(t)$  resides in a space  $\mathcal{W}_0$  spanned by the next subspace of the scaling function  $\mathcal{V}_1$ , i.e.,  $\mathcal{W}_0 \subset \mathcal{V}_1$ . Hence, in the same context as in the previous subsection,  $\psi(t)$  can be represented by a weighted sum of the shifted scaling function  $\phi(2t)$  by

$$\psi(t) = \sum_{n=-\infty}^{\infty} q_n \sqrt{2} \phi(2t-n), \quad n \in \mathbf{Z}. \quad (14.29)$$

The expansion coefficients  $q_n$  are called **wavelet coefficients** and are given by

$$q_n = (-1)^{n-1} p_{-n-1} \quad (14.30)$$

as stated below.

♠ **Theorem:**

If  $\{\mathcal{V}_m\}$  is a multiresolution analysis with the scaling function  $\phi(t)$ , the mother wavelet  $\psi(t)$  is given by

$$\psi(t) = \sqrt{2} \sum_{n=-\infty}^{\infty} (-1)^{n-1} p_{-n-1} \phi(2t-n), \quad n \in \mathbf{Z}, \quad (14.31)$$

where  $p_n$  is the scaling function coefficient of  $\phi(t)$ .

Remember that  $p_n$  in (14.31) is uniquely determined by the function form of the scaling function  $\phi(t)$ ; See (14.28). Thus the above theorem states that the mother wavelet  $\psi(t)$  is obtained once the scaling function  $\phi(t)$  of a given multiresolution analysis is specified.

*Remark.* The relation  $q_n = (-1)^n p_{1-n}$  employed in equation (14.31) is one possible choice for constructing the mother wavelet  $\psi(t)$  from the father wavelet  $\phi(t)$ . In fact, there are alternative choices such as

$$q_n = (-1)^n p_{1-n}$$

or

$$q_n = (-1)^{n-1} p_{2N-1-n}$$

with certain  $N \in \mathbf{Z}$ . Hence, the mother wavelet  $\psi(t)$  associated with a given multiresolution analysis is not unique. In practice, however, any preceding definition of  $q_n$  can be used to obtain a mother wavelet  $\psi(t)$  because it leads to an orthonormal basis for the space  $\mathcal{W}_0$ .

The proof of equation (14.31) requires the following two lemmas:

♠ **Lemma 1:**

The Fourier transform  $\Phi(\omega)$  of the scaling function  $\phi(t)$  satisfies

$$\Phi(\omega) = M\left(\frac{\omega}{2}\right) \Phi\left(\frac{\omega}{2}\right),$$

where  $M(\omega)$  is the **generating function of the multiresolution analysis** defined by

$$M(\omega) = \frac{1}{\sqrt{2}} \sum_{n=-\infty}^{\infty} p_n e^{-in\omega} \quad (14.32)$$

with the scaling function coefficient  $p_n$  of  $\phi(t)$ .

♠ **Lemma 2:**

The Fourier transform  $F(\omega)$  of any function  $f \in \mathcal{W}_0$  can be expressed by

$$F(\omega) = V(\omega) e^{i\omega/2} M^*\left(\frac{\omega}{2} + \pi\right) \Phi\left(\frac{\omega}{2}\right), \quad (14.33)$$

where  $V(\omega)$  is a  $2\pi$ -periodic function, i.e.,  $V(\omega) = V(\omega + 2\pi)$ .

We should keep in mind that  $V(\omega)$  is the only term on the right-hand side of (14.33) that depends on  $f(t)$ ; the remainder term  $e^{i\omega/2} M^*[(\omega/2) + \pi] \Phi(\omega/2)$  is independent of  $f(t)$ . The proofs of the two lemmas are outlined in Exercises **3** and **4**. Now we turn to a proof of equation (14.31) for the construction of the mother wavelet  $\psi(t)$  from the scaling function  $\phi(t)$ .

*Proof (of Theorem):* Since the mother wavelet  $\psi(t)$  gives an orthonormal basis  $\{\psi(t-n), n \in \mathbf{Z}\}$  for the space  $\mathcal{W}_0$ , any function  $f \in \mathcal{W}_0$  can be expressed by

$$f(t) = \sum_{n=-\infty}^{\infty} h_n \psi(t-n)$$

with appropriate coefficients  $h_n$ . Its Fourier transform  $F(\omega)$  reads

$$F(\omega) = \left( \sum_{n=-\infty}^{\infty} h_n e^{-in\omega} \right) \Psi(\omega),$$

where the sum in parentheses is  $2\pi$ -periodic. Comparing this with (14.33), we obtain

$$\Psi(\omega) = e^{i\omega/2} M^* \left( \frac{\omega}{2} + \pi \right) \Phi \left( \frac{\omega}{2} \right). \quad (14.34)$$

Substituting expression (14.32) into (14.34) yields

$$\begin{aligned} \Psi(\omega) &= \frac{e^{i\omega/2}}{\sqrt{2}} \sum_{n=-\infty}^{\infty} p_n e^{in[(\omega/2)+\pi]} \Phi \left( \frac{\omega}{2} \right) \\ &= \frac{1}{\sqrt{2}} \sum_{n=-\infty}^{\infty} p_n e^{in\pi} e^{i(n+1)(\omega/2)} \Phi \left( \frac{\omega}{2} \right) \\ &= \frac{1}{\sqrt{2}} \sum_{k=-\infty}^{\infty} p_{-k-1} (-1)^{k-1} e^{-ik\omega/2} \Phi \left( \frac{\omega}{2} \right) \quad [k \equiv -n-1]. \end{aligned}$$

Take the inverse Fourier transform of the both sides to find

$$\begin{aligned} \psi(t) &= \frac{1}{\sqrt{2}} \sum_{k=-\infty}^{\infty} p_{-k-1} (-1)^{k-1} \int_{-\infty}^{\infty} e^{-ik\omega/2} e^{i\omega t} \Phi \left( \frac{\omega}{2} \right) d\omega \\ &= \frac{2}{\sqrt{2}} \sum_{k=-\infty}^{\infty} p_{-k-1} (-1)^{k-1} \int_{-\infty}^{\infty} e^{i\omega'(2t-k)} \Phi(\omega') d\omega' \quad [\omega' \equiv \omega/2] \\ &= \sqrt{2} \sum_{k=-\infty}^{\infty} p_{-k-1} (-1)^{k-1} \phi(2t-k). \end{aligned}$$

This is our desired result (14.31). ♣

### 14.2.8 Multiresolution Representation

Through the discussions thus far, we have obtained an orthonormal basis consisting of scaling functions  $\phi_{j,k}(t)$  and wavelets  $\psi_{j,k}(t)$  that span all of  $L^2(\mathbf{R})$ . Since

$$L^2 = \mathcal{V}_{j_0} \oplus \mathcal{W}_{j_0} \oplus \mathcal{W}_{j_0+1} \oplus \cdots,$$

any function  $x(t) \in L^2(\mathbf{R})$  can be expanded, e.g.,

$$x(t) = \sum_{k=-\infty}^{\infty} S_{j_0,k} \phi_{j_0,k}(t) + \sum_{k=-\infty}^{\infty} \sum_{j=j_0}^{\infty} T_{j,k} \psi_{j,k}(t). \quad (14.35)$$

Here, the initial scale  $j_0$  could be zero or another integer or negative infinity as in (14.13), where no scaling functions are used. The coefficients  $T_{j,k}$  are

identified with the discrete wavelet transform given in (14.11). Often  $T_{j,k}$  in (14.35) is called the **wavelet coefficient** and  $S_{j,k}$  is called the **approximation coefficient**.

The representation (14.35) can be simplified by using the following notation. We denote the first summation on the right-hand side of (14.35) by

$$x_{j_0}(t) = \sum_{k=-\infty}^{\infty} S_{j_0,k} \phi_{j_0,k}(t). \quad (14.36)$$

Equation (14.36) is called the **continuous approximation** of the signal  $x(t)$  at scale  $j_0$ . Observe that the continuous approximation approaches  $x(t)$  in the limit of  $j_0 \rightarrow \infty$ , since in this case  $L^2 = \mathcal{V}_\infty$ . In addition, we introduce the notation

$$z_j(t) = \sum_{k=-\infty}^{\infty} T_{j,k} \psi_{j,k}(t), \quad (14.37)$$

where  $z_j(t)$  is known as the **signal detail** at scale  $j$ . With these conventions, we can write (14.35) as

$$x(t) = x_{j_0}(t) + \sum_{j=j_0}^{\infty} z_j(t). \quad (14.38)$$

Equation (14.38) says that the original continuous signal  $x(t)$  is expressed as a combination of its continuous approximation  $x_{j_0}$  at an arbitrary scale index  $j_0$  added to a succession of signal details  $z_j(t)$  from scales  $j_0$  up to infinity.

Also noteworthy is the fact that due to the nested relation of  $\mathcal{V}_{j+1} = \mathcal{V}_j \oplus W_j$ , we can write

$$x_{j+1}(t) = x_j(t) + z_j(t). \quad (14.39)$$

This indicates that if we add the signal detail at an arbitrary scale (index  $j$ ) to the continuous approximation at the same scale, we get the signal approximation at an increased resolution (i.e., at a smaller scale, index  $j+1$ ). The important relation (14.39) between continuous approximations  $x_j(t)$  and signal details  $z_j(t)$  is called a **multiresolution representation**.

## Exercises

1. Verify the orthonormality of the Haar discrete wavelet  $\psi_{m,n}(t)$  defined by  $\psi_{m,n}(t) = 2^{m/2} \psi(2^m t - n)$ , where

$$\psi(t) = \begin{cases} 1 & 0 \leq t < 1/2, \\ -1 & 1/2 \leq t < 1, \\ 0 & \text{otherwise.} \end{cases}$$

**Solution:** First we note that the norm of  $\psi_{m,n}(x)$  is unity:

$$\begin{aligned}\int_{-\infty}^{\infty} \psi_{m,n}(t)^2 dt &= 2^{-m} \int_{-\infty}^{\infty} [\psi_{m,n}(2^{-m}t - n)]^2 dt \\ &= 2^{-m} \cdot 2^m \int_{-\infty}^{\infty} \psi_{m,n}(u)^2 du = 1.\end{aligned}$$

Thus, we obtain

$$\begin{aligned}I &\equiv \int_{-\infty}^{\infty} \psi_{m,n}(t)\psi_{k,\ell}(t)dt = \int_{-\infty}^{\infty} 2^{-m/2}\psi(2^{-m}t - n)2^{-k/2}\psi(2^{-k}t - \ell)dt \\ &= 2^{-m/2} \cdot 2^m \int_{-\infty}^{\infty} \psi(u)2^{-k/2}\psi[2^{m-k}(u + n) - \ell]dt.\end{aligned}\quad (14.40)$$

If  $m = k$ , the integral in the last line in (14.40) reads

$$\int_{-\infty}^{\infty} \psi(u)\psi(u + n - \ell)dt = \delta_{0,n-\ell} = \delta_{n,\ell},$$

since  $\psi(u) \neq 0$  in  $0 \leq u \leq 1$  and  $\psi(u + n - \ell) \neq 0$  in  $\ell - n \leq u \leq \ell - n + 1$ , so that these intervals are disjoint unless  $n = \ell$ . Owing to symmetry if  $m \neq k$ , it suffices to look at the case of  $m > k$ . Set  $r = m - k \neq 0$  in (14.40) to obtain

$$\begin{aligned}I &= 2^{r/2} \int_{-\infty}^{\infty} \psi(u)\psi(2^r v + s)du \\ &= 2^{r/2} \left[ \int_0^{1/2} \psi(2^r v + s)du - \int_{1/2}^1 \psi(2^r v + s)du \right],\end{aligned}$$

which can be simplified as

$$I = \int_s^a \psi(x)dx - \int_a^b \psi(x)dx = 0, \quad (14.41)$$

where  $2^r u + s = x$ ,  $a = s + 2^{r-1}$ ,  $b = s + 2^r$ . Observe that  $[s, a]$  contains the interval  $[0, 1]$  of the Haar wavelet  $\psi(t)$ , which implies that the first integral in (14.41) vanishes. Similarly, the second integral equals zero. We thus conclude that

$$I = \int_{-\infty}^{\infty} \psi_{m,n}(t)\psi_{k,\ell}dt = \delta_{m,k}\delta_{n,\ell},$$

which means that the Haar discrete wavelet  $\psi_{m,n}(t)$  is orthonormal.



2. Let  $\phi \in L^2(\mathbf{R})$  and  $\Phi(\omega)$  be the Fourier transform of  $\phi(t)$ . Prove that the system  $\{\phi_{0,n} \equiv \phi(t-n), \quad n \in \mathbf{Z}\}$  is orthonormal if and only if  $\sum_{k=-\infty}^{\infty} |\Phi(\omega + 2k\pi)|^2 = 1$  almost everywhere.

**Solution:** It is obvious that the Fourier transform of  $\phi_{0,n}(t)$  reads  $\Phi_{0,n}(\omega) = e^{-in\omega}\Phi(\omega)$ . In view of the Parseval identity for the wavelet transform (14.8), we have

$$\begin{aligned} \int_{-\infty}^{\infty} \phi_{0,n}(t)\phi_{0,m}(t)dt &= \int_{-\infty}^{\infty} \phi_{0,0}(t)\phi_{0,m-n}(t)dt \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \Phi_{0,0}(\omega)\Phi_{0,m-n}(\omega)d\omega = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-i(m-n)\omega} [\Phi_{0,0}(\omega)]^2 d\omega \\ &= \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \int_{2\pi k}^{2\pi(k+1)} e^{-i(m-n)\omega} [\Phi_{0,0}(\omega)]^2 d\omega \\ &= \frac{1}{2\pi} \int_0^{2\pi} e^{-i(m-n)\omega} \sum_{k=-\infty}^{\infty} [\Phi_{0,0}(\omega)]^2 d\omega. \end{aligned}$$

It thus follows from the completeness of  $\{e^{-in\omega}, \quad n \in \mathbf{Z}\}$  in  $L^2(0, 2\pi)$  that  $\int_{-\infty}^{\infty} \phi_{0,n}(t)\phi_{0,m}(t)dt = 0$  if and only if

$$\sum_{k=-\infty}^{\infty} [\Phi_{0,0}(\omega)]^2 = 1 \quad \text{almost everywhere.} \quad \clubsuit$$

3. Let  $\Phi(\omega)$  be the Fourier transform of the scaling function  $\phi(t)$  and let  $p_n$  be its scaling function coefficient. Prove that

$$\Phi(\omega) = M\left(\frac{\omega}{2}\right)\Phi\left(\frac{\omega}{2}\right) \quad \text{with} \quad M(\omega) = \frac{1}{\sqrt{2}} \sum_{n=-\infty}^{\infty} p_n e^{-in\omega}. \quad (14.42)$$

**Solution:** Since  $\phi(t) = \sqrt{2} \sum_{n=-\infty}^{\infty} p_n \phi(2t-n)$ , we have

$$\begin{aligned} \Phi(\omega) &= \sqrt{2} \sum_{n=-\infty}^{\infty} p_n \int_{-\infty}^{\infty} \phi(2t-n)e^{-i\omega t} dt \\ &= \sqrt{2} \sum_{n=-\infty}^{\infty} p_n \int_{-\infty}^{\infty} \phi(t')e^{-i\omega(t'+n)/2} dt' \quad (t' \equiv 2t-n) \\ &= \frac{1}{\sqrt{2}} \sum_{n=-\infty}^{\infty} p_n e^{-in\omega/2} \Phi\left(\frac{\omega}{2}\right) = M\left(\frac{\omega}{2}\right)\Phi\left(\frac{\omega}{2}\right). \quad \clubsuit \end{aligned}$$

4. Let  $f(t)$  be a function  $f \in \mathcal{W}_0 = \mathcal{V}_1 \setminus \mathcal{V}_0$  for a given multiresolution analysis  $\{\mathcal{V}_j\}$ . Prove that its Fourier transform  $F(\omega)$  necessarily takes the form

$$F(w) = V(\omega)e^{i\omega/2}M^*\left(\frac{\omega}{2} + \pi\right)\Phi\left(\frac{\omega}{2}\right), \quad (14.43)$$

where  $V(\omega) = V(\omega + 2\pi)$ .

**Solution:** Since  $f \in \mathcal{W}_0$  and  $\mathcal{V}_1 = \mathcal{V}_0 \oplus \mathcal{W}_0$ , it follows that  $f \in \mathcal{V}_1$  and is orthogonal to  $\mathcal{V}_0$ . Hence, we can write  $f(t) = \sqrt{2} \sum_{n=-\infty}^{\infty} c_n \phi_{1,n}(t) = \sqrt{2} \sum_{n=-\infty}^{\infty} c_n \phi(2t - n)$ , where  $c_n = \int_{-\infty}^{\infty} f(t) \phi_{1,n}(t) dt$ . Take the Fourier transform of both sides to obtain

$$F(w) = M_f\left(\frac{\omega}{2}\right)\Phi\left(\frac{\omega}{2}\right) \quad \text{with} \quad M_f(\omega) \equiv \frac{1}{\sqrt{2}} \sum_{n=-\infty}^{\infty} c_n e^{-in\omega}. \quad (14.44)$$

Evidently,  $M_f(\omega)$  is a  $2\pi$ -periodic function belonging to  $L^2(0, 2\pi)$ . Since  $f$  is orthogonal to  $\mathcal{V}_0$ , we have  $\int_{-\infty}^{\infty} F(w)\Phi^*(\omega)e^{in\omega}d\omega = 0$ , so

$$\int_{-\infty}^{\infty} \left[ \sum_{k=-\infty}^{\infty} F(w + 2k\pi)\Phi^*(\omega + 2k\pi) \right] e^{in\omega} d\omega = 0.$$

Consequently,  $\sum_{k=-\infty}^{\infty} F(w + 2k\pi)\Phi^*(\omega + 2k\pi) = 0$ . Substituting (14.42) and (14.44) into this result, we obtain

$$\sum_{k=-\infty}^{\infty} M_f\left(\frac{\omega}{2} + k\pi\right) M^*\left(\frac{\omega}{2} + k\pi\right) \left| \Phi\left(\frac{\omega}{2} + k\pi\right) \right|^2 = 0.$$

Meanwhile we denote  $M_f(\omega)M^*(\omega)$  and  $|\Phi(\omega)|^2$  by  $M_2(\omega)$  and  $\Phi_2(\omega)$ , respectively. By splitting the sum into even and odd integers  $k$  and then employing the  $2\pi$ -periodicity of  $M(\omega)$  and  $M_f(\omega)$  [and thus  $M_2(\omega)$ ], we have

$$\begin{aligned} 0 &= \sum_{k=-\infty}^{\infty} M_2\left(\frac{\omega}{2} + 2k\pi\right) \Phi_2\left(\frac{\omega}{2} + 2k\pi\right) \\ &\quad + \sum_{k=-\infty}^{\infty} M_2\left[\frac{\omega}{2} + (2k+1)\pi\right] \Phi_2\left[\frac{\omega}{2} + (2k+1)\pi\right] \\ &= M_2\left(\frac{\omega}{2}\right) \sum_{k=-\infty}^{\infty} \Phi_2\left(\frac{\omega}{2} + 2k\pi\right) \\ &\quad + M_2\left(\frac{\omega}{2} + \pi\right) \sum_{k=-\infty}^{\infty} \Phi_2\left[\frac{\omega}{2} + (2k+1)\pi\right] \\ &= M_2\left(\frac{\omega}{2}\right) + M_2\left(\frac{\omega}{2} + \pi\right), \end{aligned} \quad (14.45)$$

where we used the orthonormality condition with respect to the set of scaling functions  $\{\phi_{0,k}(t)\}$ . Finally, replacing  $\omega$  in the last line in (14.45) by  $2\omega$  gives

$$\begin{vmatrix} M_f(\omega) & M^*(\omega + \pi) \\ -M_f(\omega + \pi) & M^*(\omega) \end{vmatrix} = 0, \quad (14.46)$$

which indicates the linear dependence of two vectors:  $[M_f(\omega), -M_f(\omega + \pi)]$  and  $[M^*(\omega + \pi), M^*(\omega)]$ . Hence, there exists a function  $\lambda(\omega)$  such that

$$M_f(\omega) = \lambda(\omega)M^*(\omega + \pi). \quad (14.47)$$

Since both  $M$  and  $M_f$  are  $2\pi$  periodic, so is  $\lambda$ . Further, substituting (14.47) into (14.46) yields

$$\lambda(\omega) + \lambda(\omega + \pi) = 0, \quad (14.48)$$

which means that there exists a function  $V(\omega)$  such that

$$\lambda(\omega) = e^{i\omega}V(\omega) \quad \text{and} \quad V(\omega) = V(\omega + 2\pi).$$

Eventually, the results (14.44), (14.47), and (14.48) lead to the desired representation (14.43). ♣

## 14.3 Fast Wavelet Transformation

### 14.3.1 Generalized Two-Scale Relations

We know that a signal  $x(t) \in L^2(\mathbf{R})$  can be represented in terms of the continuous approximation  $S_{m,n}$  and the discrete wavelet transform  $T_{m,n}$  by

$$x(t) = \sum_{n=-\infty}^{\infty} S_{m_0,n} \phi_{m_0,n}(t) + \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} T_{m,n} \psi_{m,n}(t),$$

where

$$\phi_{m,n}(t) = 2^{m/2} \phi(2^m t - n) \quad \text{and} \quad \psi_{m,n}(t) = 2^{m/2} \psi(2^m t - n). \quad (14.49)$$

[See (14.24) and (14.26).] In principle, both expansion coefficients  $S_{m,n}$  and  $T_{m,n}$  can be computed through the convolution integral defined by

$$S_{m,n} = \int_{-\infty}^{\infty} x(t) \phi_{m,n}(t) dt \quad \text{and} \quad T_{m,n} = \int_{-\infty}^{\infty} x(t) \psi_{m,n}(t) dt. \quad (14.50)$$

Actual computations of these integrals are very time-consuming. However, there is an efficient method for computing  $S_{m,n}$  and  $T_{m,n}$  at all  $m$ , known as

the **fast wavelet transform**. This sophisticated method is based on recursive equations for  $S_{m,n}$  and  $T_{m,n}$  and thus is markedly suitable for numerical computations of wavelet analyses.

To proceed with the argument, we need some preliminary results. We know that the father wavelet  $\phi(t)$  and the mother wavelet  $\psi(t)$  can be described by a linear combination of contracted and shifted versions of  $\phi(t)$  as follows:

$$\phi(t) = \sqrt{2} \sum_{k=-\infty}^{\infty} p_k \phi(2t - k) \quad \text{and} \quad \psi(t) = \sum_{k=-\infty}^{\infty} (-1)^k p_{1-k} \phi(2t - k),$$

where  $p_n$  is the scaling function coefficient of  $\phi(t)$ . For convenience, we use an alternative definition  $q_n = (-1)^n p_{1-n}$  of the wavelet coefficient  $q_n$  instead of the one used in (14.30). These facts immediately result in

$$\begin{aligned} \phi_{m,n}(t) &= 2^{m/2} \phi(2^m t - n) = 2^{m/2} \sum_{k=-\infty}^{\infty} p_k \phi[2(2^m t - n) - k] \\ &= 2^{m/2} \sum_{k=-\infty}^{\infty} p_k 2^{-(m+1)/2} \phi_{m+1,2n+k}(t) \\ &= 2^{-1/2} \sum_{k=-\infty}^{\infty} p_k \phi_{m+1,2n+k}(t), \end{aligned} \tag{14.51}$$

and similarly,

$$\psi_{m,n}(t) = 2^{-1/2} \sum_{k=-\infty}^{\infty} q_k \phi_{m+1,2n+k}(t). \tag{14.52}$$

The expressions (14.51) and (14.52) are generalizations of (14.27) and (14.31) applicable for  $\phi(t)$  and  $\psi(t)$ .

#### ♠ Generalized two-scale relations:

Given a multiresolution analysis,  $\phi_{m,n}(t)$  and  $\psi_{m,n}(t)$  are obtained from the set of functions  $\{\phi_{m+1,2n+k}(t); -\infty < k < \infty\}$  by

$$\begin{aligned} \phi_{m,n}(t) &= 2^{-1/2} \sum_{k=-\infty}^{\infty} p_k \phi_{m+1,2n+k}(t), \\ \psi_{m,n}(t) &= 2^{-1/2} \sum_{k=-\infty}^{\infty} q_k \phi_{m+1,2n+k}(t). \end{aligned}$$

### 14.3.2 Decomposition Algorithm

The fast wavelet transform consists of two main parts, called, respectively, the **decomposition algorithm** and the **reconstruction algorithm**, each of which gives a recursive relation between approximation coefficients  $S_{m,n}$  and wavelet coefficients  $T_{m,n}$  at neighboring scales. This subsection focuses on the former algorithm and in the next subsection deals with the latter.

*Remark.* In the literature about the fast wavelet transform, all of the terms below mean the same thing:

- discrete wavelet transform
- decomposition/reconstruction algorithm
- fast orthogonal wave transform
- multiresolution algorithm
- pyramid algorithm
- tree algorithm

The decomposition algorithm enables us to obtain  $S_{m,n}$  and  $T_{m,n}$  at all  $m$  smaller than a prescribed scale  $m_0$ , once  $S_{m_0,n}$  is given. To attain our objective, we first derive a recursive formula for  $S_{m,n}$  at two different scales, i.e.,  $S_{m,n}$  and  $S_{m+1,n}$ . From the expansion (14.49) and from the orthonormality of  $\phi_{m,n}(t)$ , it follows that

$$S_{m,n} = \int_{-\infty}^{\infty} x(t) \phi_{m,n}(t) dt.$$

Using the generalized two-scale relation (14.51), we can write

$$\begin{aligned} S_{m,n} &= \int_{-\infty}^{\infty} x(t) \left[ \frac{1}{\sqrt{2}} \sum_{k=-\infty}^{\infty} p_k \phi_{m+1,2n+k}(t) \right] dt \\ &= \frac{1}{\sqrt{2}} \sum_{k=-\infty}^{\infty} p_k \left[ \int_{-\infty}^{\infty} x(t) \phi_{m+1,2n+k}(t) dt \right] \\ &= \frac{1}{\sqrt{2}} \sum_{k=-\infty}^{\infty} p_k S_{m+1,2n+k}. \end{aligned}$$

Replacing the summation index  $k$  with  $k - 2n$ , we obtain

$$S_{m,n} = \frac{1}{\sqrt{2}} \sum_{k=-\infty}^{\infty} p_{k-2n} S_{m+1,k}, \quad (14.53)$$

which provides the approximation coefficients  $S_{m,n}$  from  $S_{m+1,n}$ .

Similarly the wavelet coefficients  $T_{m,n}$  can be found from the approximation coefficients at the previous scale:

$$T_{m,n} = \frac{1}{\sqrt{2}} \sum_{k=-\infty}^{\infty} q_k S_{m+1,2n+k} = \frac{1}{\sqrt{2}} \sum_{k=-\infty}^{\infty} q_{k-2n} S_{m+1,k}. \quad (14.54)$$

As a consequence, if we know the approximation coefficients  $S_{m_0,k}$  at a specific scale  $m_0$  then, through repeated application of (14.53) and (14.54), we can generate  $S_{m,n}$  and  $T_{m,n}$  at all  $m < m_0$ . This procedure, called the **decomposition algorithm**, which is based on (14.53) and (14.54) is the first half of the fast wavelet transform that allows us to compute the wavelet coefficients efficiently, rather than computing them laboriously from the convolution of (14.50).

### 14.3.3 Reconstruction Algorithm

We can go in the opposite direction and reconstruct  $S_{m+1,n}$  from  $S_{m,n}$  and  $T_{m,n}$ . We already know from (14.39) that  $x_{m+1}(t) = x_m(t) + z_m(t)$ , and we can expand this as

$$x_{m+1}(t) = \sum_{n=-\infty}^{\infty} S_{m,n} \phi_{m,n}(t) + \sum_{n=-\infty}^{\infty} T_{m,n} \psi_{m,n}(t).$$

Furthermore, using (14.51) and (14.52), we can expand this equation in terms of the scaling function at the previous scale:

$$\begin{aligned} x_{m+1}(t) &= \sum_{n=-\infty}^{\infty} S_{m,n} \frac{1}{\sqrt{2}} \sum_{k=-\infty}^{\infty} p_k \phi_{m+1,2n+k}(t) \\ &\quad + \sum_{n=-\infty}^{\infty} T_{m,n} \frac{1}{\sqrt{2}} \sum_{k=-\infty}^{\infty} q_k \phi_{m+1,2n+k}(t). \end{aligned}$$

Rearranging the summation indices, we get

$$\begin{aligned} x_{m+1}(t) &= \sum_{n=-\infty}^{\infty} S_{m,n} \frac{1}{\sqrt{2}} \sum_{k=-\infty}^{\infty} p_{k-2n} \phi_{m+1,k}(t) \\ &\quad + \sum_{n=-\infty}^{\infty} T_{m,n} \frac{1}{\sqrt{2}} \sum_{k=-\infty}^{\infty} q_{k-2n} \phi_{m+1,k}(t). \end{aligned} \quad (14.55)$$

We also know that we can expand  $x_{m+1}(t)$  in terms of the approximation coefficients at scale  $m+1$ , i.e.,

$$x_{m+1}(t) = \sum_{k=-\infty}^{\infty} S_{m+1,k} \phi_{m+1,k}(t). \quad (14.56)$$

Equating the coefficients in (14.56) with (14.55) yields the reconstruction algorithm:

$$S_{m+1,n} = \frac{1}{\sqrt{2}} \sum_{k=-\infty}^{\infty} p_{n-2k} S_{m,k} + \frac{1}{\sqrt{2}} \sum_{k=-\infty}^{\infty} q_{n-2k} T_{m,k},$$

where we have swapped the indices  $k$  and  $n$ . Hence, at the scale  $m + 1$ , the approximation coefficients  $S_{m+1,n}$  can be found in terms of a combination of  $S_{m,n}$  and  $T_{m,n}$  at the next scale,  $m$ . The reconstruction algorithm is the second half of the fast wavelet transform.

## Differential Equations



## Ordinary Differential Equations

**Abstract** The main objective of this chapter is to ensure that the reader understands the “existence theorem” (Sect. 15.2.3) and the “unique theorem” (Sect. 15.2.4) for a first-order ordinary differential equation. These theorems prove the existence and uniqueness of a solution of the differential equation and delineate the conditions that should be satisfied by the functions that are to be differentiated.

### 15.1 Concepts of Solutions

#### 15.1.1 Definition of Ordinary Differential Equations

Many physical laws are often formulated as **ordinary differential equations** (ODEs) whose unknowns are functions of a single variable. Below are basic notation and several important theorems that are used throughout this chapter. We start with the formal definition of ODEs.

♠ **Ordinary differential equations:**

An ordinary differential equation of order  $n$  is an equation

$$F \left[ x, y(x), y'(x), \dots, y^{(n)}(x) \right] = 0 \quad (15.1)$$

that is satisfied by the function  $y(x)$  and its derivatives  $y'(x), y''(x), \dots, y^{(n)}(x)$  with respect to a single independent variable  $x$ .

Here, the **order** of a differential equation means the largest positive integer  $n$  for which an  $n$ th derivative appears in equation (15.1). For instance, a general form of the first-order differential equations is given by

$$F \left[ x, y(x), y'(x) \right] = 0, \quad (15.2)$$

where  $F$  is a single-valued function on its arguments in some domain  $D$ . Hereafter we restrict our attention to the case where  $x$  is a real number.

*Remark.*

1. An ODE (15.1) is called a **linear ODE** if it is linear in the unknown function  $y(x)$  and in all its derivatives; otherwise, it is **nonlinear**.
2. A linear ODE of order  $n$  is said to be **homogeneous** if it is of the form  $a_n(x)y^{(n)} + a_{n-1}(x)y^{(n-1)} + \dots + a_1(x)y' + a_0(x)y = 0$ , where there is no term that contains a function of  $x$  alone.
3. The term **homogeneous** may have a totally different meaning specifically when a linear ODE is first order, which occurs if the ODE is written in the form

$$\frac{dy}{dx} = F\left(\frac{y}{x}\right). \quad (15.3)$$

Such equations can be solved in closed form by a change of variables  $u = y/x$ , which transforms the equation into the separable equation

$$\frac{dx}{x} = \frac{du}{F(u) - u}. \quad (15.4)$$

### 15.1.2 Explicit Solution

Let  $y = \varphi(x)$  define  $y$  as a function of  $x$  on an interval  $I = (a, b)$ . We say that the function  $\varphi(x)$  is an **explicit solution** or a simple **solution** of the ODE (15.1) if it satisfies the equation for *every*  $x$  in  $I$ . In mathematical symbols, this definition reads as follows:

♠ **Explicit solution of an ODE:**

A function  $y = \varphi(x)$  defined on an interval  $I$  is a solution of the ODE (15.1) if

$$F\left[x, \varphi(x), \varphi'(x), \dots, \varphi^{(n)}(x)\right] = 0$$

for *every*  $x$  in  $I$ .

Note that a real function should be a correspondence between two sets of real numbers. In this context, if an equation involving  $x$  and  $y$  does not define a real function, then it is not a solution of any ODE even if the equation formally satisfies the ODE. For example, the equation

$$y = \sqrt{-(1+x^2)} \quad (15.5)$$

does not define a real function; therefore, it is not a solution of the ODE

$$x + yy' = 0 \quad (15.6)$$

even though the formal substitution of (15.5) into (15.6) yields an identity.

*Examples* **1.** The function

$$y = \log x + c, \quad x > 0$$

is a solution of  $y' = 1/x$  for all  $x > 0$ .

**2.** The function

$$y = \tan x - x, \quad x \neq \frac{2n+1}{2}\pi \quad (n = 0, \pm 1, \pm 2, \dots) \quad (15.7)$$

is a solution of

$$y' = (x + y)^2. \quad (15.8)$$

In fact, the substitution of  $y$  into (15.8) gives the identity  $\tan^2 x = (x + \tan x - x)^2 = \tan^2 x$  in each of the intervals specified in (15.7).

**Remark.** Note that the ODE (15.8) is defined for all  $x$ , but its solution (15.7) is not defined for all  $x$ . Hence, the interval for which the function given by (15.7) may be a solution of (15.8) is a smaller set of the intervals in (15.7).

**3.** The function  $y = |x|$  is a solution of

$$y' = 1 \quad \text{in the interval } x > 0,$$

and is also a solution of

$$y' = -1 \quad \text{in the interval } x < 0.$$

**Remark.** Observe that the function  $y = |x|$  is defined for all  $x$ , whereas the corresponding ODEs are defined in only a restricted interval of  $x$ , in contrast to Example **2**.

### 15.1.3 Implicit Solution

It is sometimes not easy (or even impossible) to solve an equation of the form  $g(x, y) = 0$  for  $y$  in terms of  $x$ . However, whenever it can be shown that an implicit function does satisfy a given ODE on an interval  $I$ , then the relation  $g(x, y) = 0$  is called an **implicit solution** of the ODE. A formal definition is given below.

♠ **Implicit solution of an ODE:**

A relation  $g(x, y) = 0$  is an **implicit solution** of an ODE

$$F\left[x, y(x), y'(x), \dots, y^{(n)}(x)\right] = 0$$

on an interval  $I$  if:

1. There exists a function  $h(x)$  defined on  $I$  such that  $g(x, h(x)) = 0$  for every  $x$  in  $I$ .
2. If  $F\left[x, h(x), h'(x), \dots, h^{(n)}(x)\right] = 0$  for every  $x$  in  $I$ .

**Remark.** It must be cautioned that  $g(x, y) = 0$  is merely an *equation*, and it is thus never a precise solution of an ODE, as only a *function* can be a solution of an ODE. What we mean in the above definition is that the function  $h(x)$  defined by the relation  $g(x, y) = 0$  is the solution of the ODE.

*Examples* The equation

$$g(x, y) = x^2 + y^2 - 25 = 0$$

is an implicit solution of the ODE

$$F(x, y, y') = yy' + x = 0$$

on the interval  $I : -5 < x < 5$ . In fact, the function  $h(x) = \sqrt{25 - x^2}$  defined on  $I$  yields

$$F[x, h(x), h'(x)] = \sqrt{25 - x^2} \left( -\frac{x}{\sqrt{25 - x^2}} \right) + x = 0$$

for every  $x$  on  $I$ .

#### 15.1.4 General and Particular Solutions

We next observe that an ODE in general has many solutions. For example, the ODE

$$y' = e^x$$

can be solved as

$$y = e^x + c, \tag{15.9}$$

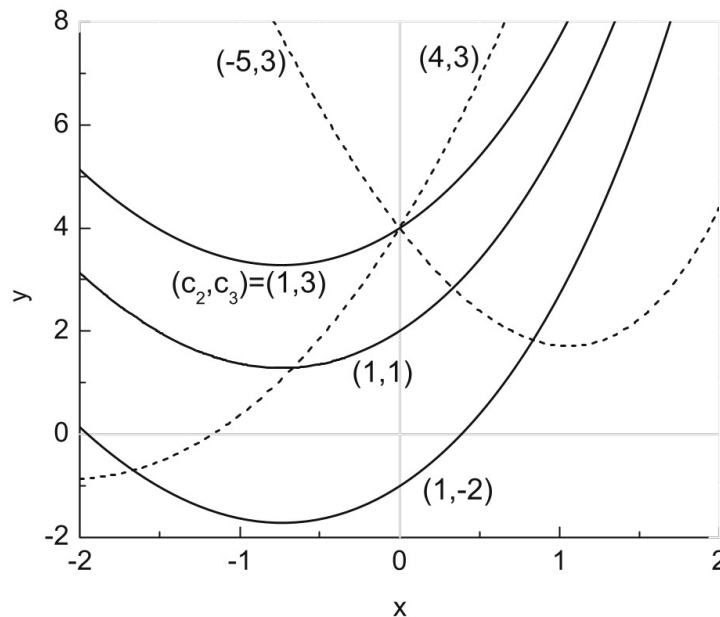
where  $c$  can take any numerical value. Similarly, if

$$y''' = e^x, \tag{15.10}$$

then its solution, obtained by integrating three times, is

$$y = e^x + c_1 x^2 + c_2 x + c_3, \tag{15.11}$$

where  $c_1, c_2, c_3$  can take on any numerical values. Note that both (15.9) and (15.11) express infinitely many solutions since, which are constants, the  $c$ 's can have infinitely many values. Figure 15.1 is a geometrical interpretation of this point. Each curve corresponds to a solution (15.11) for  $c_2 = -5, 1, 4$  and  $c_3 = -2, 1, 3$  while  $c_1 = 1$  is fixed.



**Fig. 15.1.** Family of the infinitely many solutions (15.11) of the differential equation (15.10). *Solid and dotted curves* correspond to  $c_2 = 1$  and  $c_2 = 3$ , respectively

The two examples above illustrate that solutions of an ODE may often be represented by a single equation involving an arbitrary constant  $c$ . Such a function involving an arbitrary constant is called a **general solution** (or **complete integral** or **primitive integral**) of an ODE. Geometrically, these are infinitely many curves, one for each set of values of the  $c$ 's. If we choose specific values of the  $c$ 's, we obtain what is called a **particular solution** of that ODE.

*Remark.* From the examples above, the reader might assume that

- (i) an ODE always has infinitely many solutions, or that
- (ii) a solution of an  $n$ th order ODE always contains  $n$  arbitrary constants.

However, these two conjectures are false. For instance,

- The equation  $(y'')^2 + y^2 = 0$  has only one solution  $y = 0$  that possesses *no* arbitrary constant.
- The equation  $|y'| + 1 = 0$  has *no* solution.
- The *first-order* equation  $(y' - y)(y' - 2y) = 0$  has the solution  $(y - c_1 e^x)(y - c_2 e^{2x}) = 0$  that has *two* (not *one*) arbitrary constants.

### 15.1.5 Singular Solution

Consider an ODE of the form

$$y - xy' = f(y'), \quad (15.12)$$

which is known as a **Clairaut equation**. We solve it by differentiating both sides to yield

$$y'' [f'(y') + x] = 0.$$

We thus have two possibilities. If we set  $y'' = 0$ , then  $y = ax + b$  so that substitution back into the original equation (15.12) gives  $b = f(a)$ . Thus we have a general solution:

$$y = ax + f(a),$$

where  $a$  is an arbitrary constant. On the other hand, if we set

$$f'(y') + x = 0, \quad (15.13)$$

then eliminating  $y'$  between (15.13) and the original equation gives us a solution with *no* arbitrary constant, which is known as a **singular solution**. There are various other types of singular solutions, one of which is given below.

*Examples* Suppose the Clairaut equation to be of the form

$$y = xy' + (y')^2$$

and differentiate both sides to obtain

$$y''(x + 2y') = 0.$$

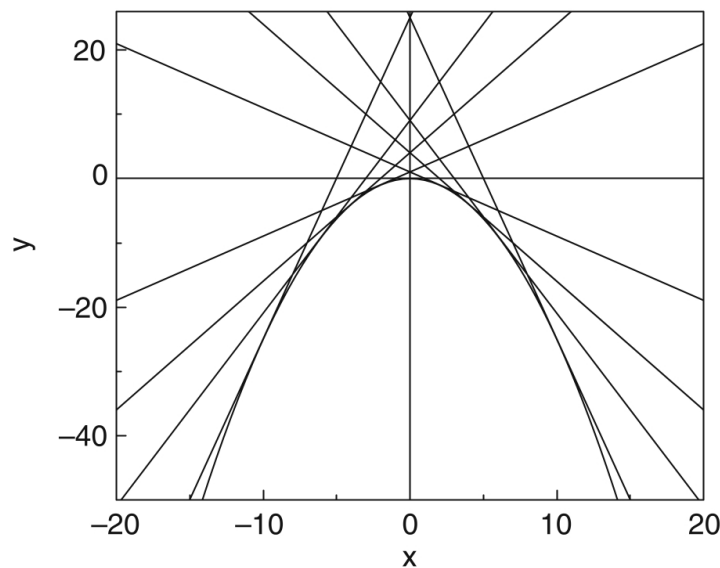
If we set  $y'' = 0$ , then the general solution reads

$$y = cx + c^2 \quad (15.14)$$

with an arbitrary constant  $c$ . However, if we choose the possibility that  $2y' + x = 0$ , then we have

$$x^2 + 4y = 0. \quad (15.15)$$

**Remark.** Geometrically, the singular solution (15.14) is an *envelope* of the family of integral curves defined by the general solution (15.15), as depicted in Fig. 15.2. The dotted parabola is the singular solution and the straight lines tangent to the parabola are the general solution.



**Fig. 15.2.** The singular solution (15.14) is an envelope of the family of **integral curves** (see Sect. 15.1.6), which are defined by the general solution (15.15)

### 15.1.6 Integral Curve and Direction Field

Before closing this section, we must emphasize the geometric significance of a solution of a first-order ODE. In many practical problems, a rough geometrical approximation to a solution may be all that is needed rather than an evaluation of its explicit functional form. Let

$$y = f(x) \text{ or } g(x, y) = 0$$

define a function of  $x$  whose derivative  $y'$  exists on an interval  $I : a < x < b$ . Then  $y'$  gives the direction of the tangent to the curve at each of these points. Therefore, finding a solution for

$$y' = F(x, y), \quad a < x < b \quad (15.16)$$

can be reduced to finding a curve on the  $(x, y)$ -plane whose slope at each of its points is given by (15.16). The relevant terminology is given below.

#### ♠ **Integral curve:**

If a curve  $y = f(x)$  [or  $g(x, y) = 0$ ] satisfies a first-order ODE (15.16) on an interval  $I$ , then the graph of this function is called an **integral curve**.

Obviously, an integral curve is the graph of a function that is a solution of a first-order ODE (15.16). Therefore, even if we cannot find an elementary function that is a solution of (15.16), we can draw a small line element at any point on the  $(x-y)$ -plane for which  $x$  is in  $I$  to represent the slope of an integral curve. If this line is short enough, the curve itself over that length resembles the line. These lines are called **line elements** and an ensemble of such lines is called a **direction field**.

## Exercises

1. Test whether the relation

$$xy^2 - e^{-y} - 1 = 0 \quad (15.17)$$

is an implicit solution of the ODE

$$(xy^2 + 2xy - 1)y' + y^2 = 0. \quad (15.18)$$

**Solution:** If we blindly differentiate both sides to yield

$$2xyy' + y^2 + e^{-y}y' = 0 \quad (2xy + e^{-y})y' + y^2 = 0$$

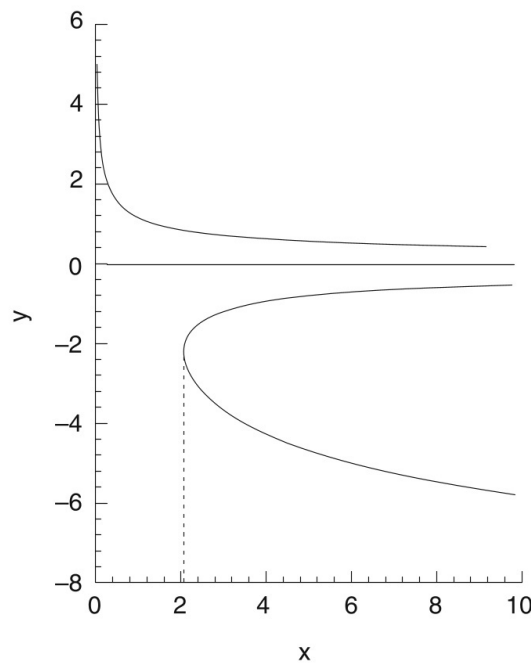
and then eliminate  $e^{-y}$  from the final result by using (15.17), we obtain the ODE (15.18). This implies the possibility that (15.17) is an implicit solution of the ODE (15.18). The remaining task is, therefore, to determine the *interval*  $I$  on which we can define such a function  $y = h(x)$  that satisfies the relation (15.17) for *every*  $x$  on  $I$ .

As a first step, we write (15.17) as

$$y = \pm \sqrt{\frac{1 + e^{-y}}{x}},$$

which says that  $y$  is defined only for  $x > 0$  since  $e^{-y}$  is always positive. Hence, the interval for which (15.17) may be a solution of (15.18) must exclude values of  $x \leq 0$ .

Next, we depict a graph of equation (15.17) on the  $(x-y)$ -plane (see Fig. 15.3). From the graph, we see that there are three choices for the function  $y = h(x)$ , each of which gives a one-to-one relation between  $x$  and  $y$ . If we choose the upper branch ( $y > 0$ ), then we can say that “(15.17) is an implicit solution of (15.18) *for all*  $x > 0$ .” If we choose either of the two lower branches (one is above the dashed line and the other is below), then we can say that “(15.17) is an implicit solution of (15.18) *only for*  $x > x_0 \simeq 2.07$ .” ♣



**Fig. 15.3.** The curve of the function (15.17)

## 15.2 Existence Theorem for the First-Order ODE

### 15.2.1 Picard Method

In this section, we consider a first-order ODE of the form

$$y'(x) = f(x, y(x)), \quad (15.19)$$

where  $f$  is some continuous function. Our main purpose is to prove that:

- (i) a wide class of equations of the form (15.19) have solutions, and
- (ii) solutions to **initial value problems**

$$y'(x) = f(x, y(x)), \quad y(x_0) = y_0$$

are unique. Statements (i) and (ii) are supported by the **existence theorem** and the **uniqueness theorem**, respectively, as is demonstrated in the subsequent subsections.

Our proof of the two theorems is based on that we call **Picard's method**, which gives solutions of an initial value problem

$$y'(x) = f(x, y(x)), \quad y(x_0) = y_0, \quad (15.20)$$

where  $f(x, y(x))$  is assumed to be continuous and real-valued in a rectangle:

$$R: |x - x_0| < a, \quad |y - y_0| < b \quad (a, b > 0). \quad (15.21)$$

The key to Picard's method is to replace the differential equation in (15.20) by the equivalent integral form,

$$y(x) = y_0 + \int_{x_0}^x f(t, y(t)) dt, \quad (15.22)$$

which is an **integral equation** because the unknown function  $y(x)$  appears in the integrand. That the integral equation (15.22) is equivalent to the original initial value problem can be checked by differentiating (15.22) on  $x$ .

**Remark.** Note that the initial condition  $y(x_0) = y_0$  is automatically included in (15.22).

We now try to solve (15.22). As a crude approximation to a solution, we take the constant function  $\varphi_0(x) = y_0$ , which clearly satisfies the initial condition

$$\varphi_0(x_0) = y_0,$$

whereas it does not satisfy (15.22) in general. Nevertheless, if we substitute the constant function into  $f(t, y(t))$  of (15.22), we have

$$\varphi_1(x) = y_0 + \int_{x_0}^x f(t, \varphi_0(t)) dt, \quad (15.23)$$

which is a closer approximation to a solution than  $\varphi_0(x)$ . By continuing the process, we have a sequence of functions  $\{\varphi_n(x)\}$ :

♠ **Successive approximation:**

Given an integral equation (15.22) with respect to  $y(x)$ , a set of functions defined by

$$\begin{aligned} \varphi_0(x) &= y_0, \\ \varphi_n(x) &= y_0 + \int_{x_0}^x f(t, \varphi_{n-1}(t)) dt. \quad (n = 1, 2, \dots) \end{aligned} \quad (15.24)$$

is called a **successive approximation** to a solution of (15.22).

We understand intuitively that taking the limit  $n \rightarrow \infty$  yields

$$\varphi_n(x) \rightarrow \varphi(x),$$

where  $\varphi(x)$  is the exact solution of the integral equation (15.22). The convergence property of the sequence  $\{\varphi_i(x)\}$  and the equivalence of the limit function  $\varphi(x)$  to the solution of (15.22) are guaranteed if the integrand  $f(x, y(x))$  satisfies several conditions as is demonstrated in Sect. 15.2.3.

In summary, we now know the following:

♠ **Picard method:**

The differential equation  $y'(x) = f(x, y(x))$  for a given initial value  $y(x_0) = y_0$  can be solved by starting with  $\varphi_0(x) = y_0$  and then computing successive approximations (15.24). The process converges to a solution of the differential equation, where  $f(x, y)$  satisfies several specific conditions given in Sect. 15.2.3.

### 15.2.2 Properties of Successive Approximations

We have previously assumed that  $f(x, y(x))$  is continuous in the rectangle  $R$  defined in (15.21). Hereafter, we further assume that  $f(x, y(x))$  is bounded on  $R$ , which means the existence of a constant  $M > 0$  such that

$$|f(x, y(x))| \leq M \quad \text{for all } (x, y) \in R.$$

In this case, the successive approximations  $\{\varphi_n(x)\}$  show both the continuity and boundedness property stated below.

♠ **Continuity of successive approximations:**

Let  $f(x, y)$  be continuous and bounded by  $|f(x, y)| \leq M$  in a rectangle

$$R : |x - x_0| < a, \quad |y - y_0| < b \quad (a, b > 0).$$

Then, the successive approximations  $\varphi_n(x)$  are continuous on the interval

$$I : |x - x_0| \leq c \equiv \min \left[ a, \frac{b}{M} \right].$$

♠ **Boundedness of successive approximations:**

Under the same conditions as above, the  $\varphi_n(x)$  satisfy the inequality

$$|\varphi_n(x) - y_0| \leq M|x - x_0|$$

for all  $x$  in  $I$ .

**Remark.** The condition  $|f(x, y)| \leq M$  has an important geometric meaning in terms of the direction field. Since  $y' = f(x, y)$ , the direction field  $y'$  is bounded as  $|y'| \leq M$ , namely,  $-M \leq y' \leq M$  for all points in  $R$ . Therefore, a solution curve  $\varphi(x)$  that passes through  $(x_0, y_0)$  must lie in the shadowed region in Fig. 15.4.

*Proof (of the continuity).* From (15.23), we have

$$|\varphi_1(x) - y_0| = \left| \int_{x_0}^x f(t, \varphi_0(t)) dt \right| \leq \int_{x_0}^x |f(t, y_0)| dt \leq M|x - x_0|, \quad (15.25)$$

since  $\varphi_0(t) = y_0$  and  $|f(x, y_0)| \leq M$ . Now we tentatively assume that the theorem is true for a function  $\varphi_n$  with  $n \geq 1$ , and then prove inductively that it is also true for  $\varphi_n$ . By hypothesis, all points  $(t, \varphi_{n-1}(t))$  for  $t$  in  $I$  are located within  $R$ . Hence, the function

$$F_{n-1}(t) = f(t, \varphi_{n-1}(t))$$

exists for  $t$  in  $I$ , which implies that

$$\varphi_n(x) = y_0 + \int_{x_0}^x F_{n-1}(t) dt$$

exists as a continuous function on  $I$ . ♣

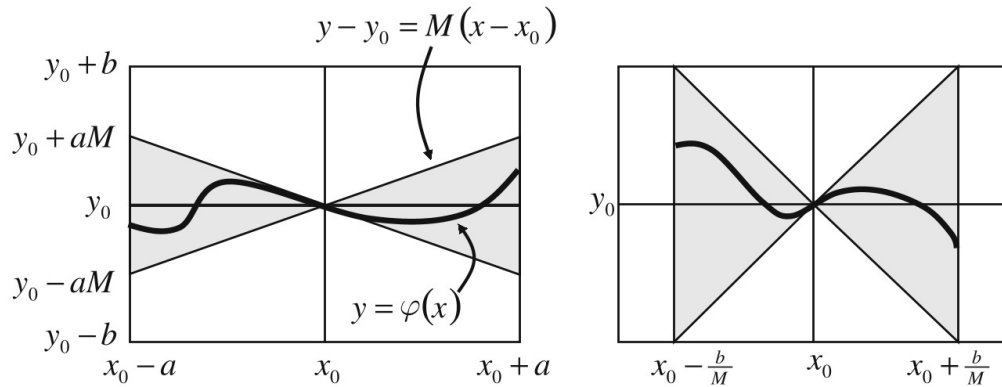
*Proof (of the boundedness).* Since by hypothesis

$$|F_{n-1}(t)| = |f(t, \varphi_{n-1}(t))| \leq M,$$

we have

$$|\varphi_n(x) - y_0| \leq \left| \int_{x_0}^x F_{n-1}(t) dt \right| \leq \int_{x_0}^x |F_{n-1}(t)| dt \leq M|x - x_0|.$$

Therefore, the boundedness of  $\varphi_n(x)$  has been proved by induction. ♣



**Fig. 15.4.** Continuity and boundedness of a solution curve  $\varphi(x)$  on the interval  $I : |x - x_0| \leq c \equiv \min[a, \frac{M}{b}]$

### 15.2.3 Existence Theorem and Lipschitz Condition

Let  $f(x, y(x))$  be a function defined for  $(x, y)$  in the rectangle  $R$  in the  $(x, y)$ -plane. We would like to verify the existence of solutions for the first-order ODEs expressed by

$$y'(x) = f(x, y(x))$$

by imposing a **Lipschitz condition**:

♠ **Lipschitz condition:**

We say that  $f(x, y(x))$  satisfies a **Lipschitz condition** on a region  $R$  if there exists a constant  $K > 0$  such that

$$|f(x, y(x)) - f(x, z(x))| \leq K |y(x) - z(x)| \quad (15.26)$$

for all  $(x, y), (x, z) \in R$ . Here the positive constant  $K$  is called the **Lipschitz constant**.

Our most important theorem is presented below.

♠ **Existence theorem:**

Suppose that

1.  $f(x, y)$  is continuous and real-valued on the rectangle  $R$ .
2.  $|f(x, y)| \leq M$  for all  $(x, y)$  in  $R$ .
3.  $f$  satisfies a Lipschitz condition with constant  $K$  in  $R$ .

Then the initial value problem

$$y'(x) = f(x, y(x)), \quad y(x_0) = y_0 \quad (15.27)$$

has at least one solution  $y(x)$  in the interval

$$I : |x - x_0| \leq c \equiv \min \left[ a, \frac{b}{M} \right].$$

*Proof* Consider the successive approximations  $\{\varphi_n(x)\}$  to a solution of the initial value problem (15.27), wherein  $f(x, y(x))$  is assumed to satisfy the Lipschitz condition (15.26). We would like to prove that **(i)** the limit function

$$\varphi(x) \equiv \lim_{n \rightarrow \infty} \varphi_n(x)$$

exists and **(ii)** that it is the solution of (15.27).

By definition of  $\varphi_n(x)$ , for  $n \geq 1$ , we obtain

$$\begin{aligned} |\varphi_{n+1}(x) - \varphi_n(x)| &\leq \left| \int_{x_0}^x [f(t, \varphi_n(t)) - f(t, \varphi_{n-1}(t))] dt \right| \\ &\leq \int_{x_0}^x |f(t, \varphi_n(t)) - f(t, \varphi_{n-1}(t))| dt \\ &\leq K \int_{x_0}^x |\varphi_n(t) - \varphi_{n-1}(t)| dt. \end{aligned} \quad (15.28)$$

Set  $n = 1$  in (15.28) and substitute it in the result (15.25) to find

$$|\varphi_2(x) - \varphi_1(x)| \leq KM \frac{|x - x_0|^2}{2!}. \quad (15.29)$$

Set  $n = 2$  in (15.28) and use the result of (15.29) in the last term in (15.28). Continuing the process, we have

$$|\varphi_n(x) - \varphi_{n-1}(x)| \leq K^{n-1} M \frac{|x - x_0|^n}{n!}. \quad (15.30)$$

Observe that the right-hand side of (15.30) is the  $n$ th term of the power series for  $e^{K|x-x_0|}$  multiplied by  $M/K$ . This implies that the infinite series

$$\varphi_0(x) + \sum_{k=1}^{\infty} [\varphi_k(x) - \varphi_{k-1}(x)] \quad (15.31)$$

is absolutely (and thus ordinary) convergent, ensuring the existence of the limit function  $\varphi(x) \equiv \lim_{n \rightarrow \infty} \varphi_n(x)$ . (See Sect. 3.2 for the convergence properties of Cauchy sequences.)

Next we prove statement (ii) above. Note that the  $n$ th partial sum of (15.31) is just  $\varphi_n(x)$  and that the infinite series (15.31) equals the limit function  $\varphi(x)$ . Hence, we have from (15.30) and (15.31) that

$$\begin{aligned} |\varphi(x) - \varphi_n(x)| &= \left| \sum_{k=n+1}^{\infty} [\varphi_k(x) - \varphi_{k-1}(x)] \right| \\ &\leq \sum_{k=n+1}^{\infty} |\varphi_k(x) - \varphi_{k-1}(x)| \leq \sum_{k=n+1}^{\infty} K^{k-1} M \frac{|x - x_0|^k}{k!} \\ &\leq \sum_{k=n+1}^{\infty} K^{k-1} M \frac{c^k}{k!} \leq \frac{M}{K} \alpha_n e^{Kc}, \end{aligned}$$

where

$$\alpha_n = \frac{(Kc)^{n+1}}{(n+1)!}.$$

Since  $\alpha_n$  is the  $n$ th term of the power series of  $e^{Kc}$ , we have  $\lim_{n \rightarrow \infty} \alpha_n = 0$ . Therefore, the series of functions  $\{\varphi_n(x)\}$  converges **uniformly** to  $\varphi(x)$  in the interval  $I : x \in [x_0 - c, x_0 + c]$ , which means that

$$\lim_{n \rightarrow \infty} f(x, \varphi_n(x)) = f(x, \varphi(x)). \quad (15.32)$$

That being so, we can write

$$\begin{aligned} \varphi(x) &= \lim_{n \rightarrow \infty} \varphi_n(x) = y_0 + \lim_{n \rightarrow \infty} \int_{x_0}^x f(t, \varphi_n(t)) dt \\ &= y_0 + \int_{x_0}^x \lim_{n \rightarrow \infty} f(t, \varphi_n(t)) dt \\ &= y_0 + \int_{x_0}^x f(t, \varphi(t)) dt. \end{aligned} \quad (15.33)$$

By differentiating on  $x$ , we have

$$\varphi(x)' = f(x, \varphi(x)), \quad \varphi(x_0) = y_0.$$

These ensure that  $\varphi(x)$  is a solution of our initial value problem (15.27). ♣

#### 15.2.4 Uniqueness Theorem

Next we examine the uniqueness of the solution  $\varphi(x)$  that we found earlier using the Picard approximation method (see Sect. 15.2.1). This is described by the theorem below.

##### ♠ Uniqueness theorem:

Let  $f(x, y)$  be continuous and satisfy the Lipschitz condition (15.26) in the rectangle  $R$ . If  $\varphi$  and  $\psi$  are two solutions of the initial value problem (15.27) in an interval  $I$  containing  $x_0$ , then  $\varphi(x) = \psi(x)$  for all  $x$  in  $I$ .

*Proof* We assume that both  $\varphi(x)$  and  $\psi(x)$  are solutions of (15.27). For  $x > x_0$ , we have from (15.33) and the Lipschitz condition (15.26) that

$$\begin{aligned} |\varphi(x) - \psi(x)| &\leq \int_{x_0}^x |f(t, \varphi(t)) - f(t, \psi(t))| dt \\ &\leq K \int_{x_0}^x |\varphi(t) - \psi(t)| dt. \end{aligned} \quad (15.34)$$

This holds in the interval  $I : x \in [x_0, x_0 + \delta]$  for arbitrary small  $\delta > 0$ . Since  $|\varphi(x) - \psi(x)|$  is continuous in  $I$ , it has a maximum at some  $x$  on  $I$ , which we label  $\mu$ . Equation (15.34) provides that

$$\mu \leq K\mu|x - x_0| \leq K\mu\delta \quad \text{for all } x \text{ in } I, \quad (15.35)$$

so we have

$$(1 - K\delta)\mu \leq 0.$$

Note that by definition  $\mu \geq 0$ . Hence, if  $K\delta < 1$ , we have  $\mu = 0$ , which says that given any Lipschitz constant  $K$ , we can find a sufficiently small  $\delta$  such that

$$\max |\varphi(x) - \psi(x)| = 0,$$

i.e.,

$$|\varphi(x) - \psi(x)| \equiv 0 \quad \text{for } x \in [x_0, x_0 + \delta].$$

Continuing this process yields the conclusion that  $|\varphi(x) - \psi(x)| \equiv 0$  for all  $x$  in  $R$ . The same holds for the case  $x < x_0$ , completing the proof. ♣

### 15.2.5 Remarks on the Two Theorems

1. The existence and uniqueness theorems only ensure the existence and uniqueness of a solution. They do not tell us whether the solution can or cannot be expressed in terms of an elementary function form or help us to find the solution.
2. Arguments for real-valued functions given thus far are straightforwardly extended to the case that  $f$  is complex-valued. In this case we must admit complex-valued solutions and  $f$  must be defined for complex  $z$ . The set of points  $z$  satisfying  $|z - z_0| \leq b$  becomes a circle with a center  $z_0$  and radius  $b$ , so domain  $R$  is no longer a rectangle.
3. The initial value problem

$$y'(x) = \sqrt{|y(x)|}, \quad y(0) = 0,$$

has two solutions,

$$y(x) \equiv 0 \quad \text{and} \quad y(x) = \begin{cases} x^2/4 & \text{if } x \geq 0, \\ -x^2/4 & \text{if } x < 0, \end{cases}$$

although  $f(x, y) = \sqrt{|y|}$  is continuous for all  $y$ . The Lipschitz condition is violated in any region that includes the line  $y = 0$  because for  $y_1 = 0$  and positive  $y_2$  we have

$$\frac{|f(x, y_2) - f(x, y_1)|}{|y_2 - y_1|} = \frac{\sqrt{y_2}}{y_2} = \frac{1}{\sqrt{y_2}} \quad (\sqrt{y_2} > 0) \quad (15.36)$$

and this can be made as large as we please by choosing  $y_2$  sufficiently small, whereas the Lipschitz condition requires that the quotient on the left-hand side of (15.36) not exceed a fixed constant  $M$ .

## Exercises

1. Using the Picard method, evaluate the successive approximation to the solution of the initial value problem

$$y'(x) = 1 + y(x)^2, \quad y(0) = 0.$$

**Solution:** Set  $x_0 = 0$ ,  $y_0 = 0$ ,  $f(x, y) = 1 + y^2$  in (15.24) to find that

$$\varphi_n(x) = \int_0^x \left\{ 1 + [\varphi_{n-1}(t)]^2 \right\} dt = x + \int_0^x [\varphi_{n-1}(t)]^2 dt.$$

Hence, we obtain

$$\begin{aligned} \varphi_1(x) &= x + \int_0^x 0 \, dt = x, & \varphi_2(x) &= x + \int_0^x t^2 \, dt = x + \frac{x^3}{3}, \\ \varphi_3(x) &= x + \int_0^x \left( t + \frac{t^3}{3} \right)^2 dt = x + \frac{x^3}{3} + \frac{2}{15}x^5 + \frac{1}{63}x^7, \text{ and so on. } \clubsuit \end{aligned}$$

**Remark.** The exact solution of the above problem can be deduced by separating variables:

$$y(x) = \tan x = x + \frac{x^3}{3} + \frac{2}{15}x^5 + \frac{17}{315}x^7 + \cdots \quad \left( -\frac{\pi}{2} < x < \frac{\pi}{2} \right).$$

The first three terms of  $\varphi_3(x)$  and of the series above are the same. The series converges only for  $|x| < \pi/2$ ; therefore, all that we can expect is that our sequence  $\varphi_1, \varphi_2, \dots$  converges to a function that is the solution of our problem for  $|x| < \pi/2$ .

2. By applying the Picard method to

$$y'(x) = xy(x), \quad y(0) = 1, \tag{15.37}$$

show that the Picard series  $\{\varphi_n(x)\}$  converges absolutely and uniformly.

**Solution:** The integral equation corresponding to (15.37) becomes

$$y(x) = 1 + \int_0^x ty(t) \, dt.$$

The iterative equation is written as  $\varphi_0(x) = 1$  and

$$\varphi_{n+1}(x) = 1 + \int_0^x t\varphi_n(t) \, dt, \quad (n = 1, 2, \dots).$$

Thus, we easily find

$$\varphi_n(x) = 1 + \frac{x^2}{2} + \frac{1}{2!} \left( \frac{x^2}{2} \right)^2 + \cdots + \frac{1}{k!} \left( \frac{x^2}{2} \right)^n.$$

The nature of the convergence is obvious for all real  $x$ , since it is a partial sum for the Taylor series of the function  $\varphi(x) = e^{x^2/2}$ .

This means that  $\varphi_n(x) \rightarrow \varphi(x)$  as  $n \rightarrow \infty$ . ♣

3. For the equation given by

$$y'(x) = 2y(x)^{1/2}, \quad y(0) = 0,$$

check the uniqueness of the solution in connection with the Lipschitz condition.

**Solution:** This equation has the two solutions  $y(x) = 0$ ,  $y(x) = 16x^2$ , although  $f(x, y) = 2(y)^{1/2}$  is continuous for all  $y$ . The Lipschitz condition (15.26) is violated in any region that includes the line  $y(x) = 0$  because for  $y_1 = 0$  and  $y_2 \neq 0$  we have

$$\frac{|f(x, y_2) - f(x, y_1)|}{|y_2 - y_1|} = \frac{\sqrt{y_2}}{y_2} = \frac{1}{\sqrt{y_2}},$$

which diverges for  $y_2 \rightarrow 0$ , exceeding a fixed constant  $K$ . ♣

## 15.3 Sturm–Liouville Problems

### 15.3.1 Sturm–Liouville Equation

ODEs encountered in physics are often classified as **Sturm–Liouville equations**:

#### ♠ Sturm–Liouville equation:

A Sturm–Liouville equation is a second-order homogeneous linear ODE of the form

$$-\frac{d}{dx} \left[ p(x) \frac{dy}{dx} \right] + q(x)y + \lambda w(x)y = 0, \quad (15.38)$$

where  $\lambda$  is a parameter and  $p$ ,  $q$ ,  $w$  are real-valued continuous functions with  $p(x) > 0$  and  $w(x) > 0$ . Here  $w(x)$  is called a **weight function**.

Using the **Sturm–Liouville operator**  $L$  defined by

$$L = \frac{1}{w(x)} \left[ -\frac{d}{dx} \left( p(x) \frac{d}{dx} \right) + q(x) \right] \quad (15.39)$$

we reduce the Sturm–Liouville equation (15.38) to the abbreviated form

$$Ly(x) = -\lambda y(x). \quad (15.40)$$

*Examples* The **Legendre equation**

$$(1 - x^2) y'' - 2xy' + n(n + 1)y = 0, \quad n \geq 1, \quad x \in [-1, 1]$$

is expressed as

$$[(1 - x^2) y']' + n(n + 1)y = 0.$$

This is in the Sturm–Liouville form of  $p = 1 - x^2$ ,  $q = 0$ ,  $w = 1$ , and  $\lambda = n(n + 1)$ .

Relevant terminology is given below.

♠ **Sturm–Liouville system:**

A **Sturm–Liouville system** consists of a Sturm–Liouville equation (15.38) on a finite closed interval  $a \leq x \leq b$ , together with two separated boundary conditions of the form

$$y(a) = \alpha y'(a) \quad \text{and} \quad y(b) = \beta y'(b)$$

with  $\alpha, \beta$  being real.

A nontrivial solution of a Sturm–Liouville system is called an **eigenfunction** and the corresponding  $\lambda$  is called an **eigenvalue**. The set of all eigenvalues of a Sturm–Liouville system is called the **spectrum** of the system.

*Examples* The Sturm–Liouville system consisting of the ODE

$$y'' + \lambda y = 0 \quad 0 \leq x \leq \pi$$

with the separated boundary conditions

$$y(0) = 0, \quad y(\pi) = 0$$

has the eigenfunction

$$y_n(x) = \sin nx$$

and the eigenvalues

$$\lambda_n = n^2. \quad (n = 1, 2, \dots).$$

### 15.3.2 Conversion into a Sturm–Liouville Equation

Mathematically, Sturm–Liouville equations represent only a small fraction of the second-order differential equations. Nevertheless, any second-order equation of the form

$$a(x)y'' + b(x)y' + c(x)y + \lambda e(x)y = 0$$

can be transformed into a Sturm–Liouville equation by multiplying the factor

$$\xi(x) = \exp \left[ \int^x \frac{b(s) - a'(s)}{a(s)} ds \right], \quad (15.41)$$

which yields a Sturm–Liouville form,

$$(\xi a y')' + \xi c y + \lambda \xi e y = 0,$$

with a nonnegative weight function  $\xi(x)a(x)$ .

*Examples* We show below that the **Hermite equation** of the form

$$y'' - 2xy' + 2\alpha y = 0 \quad (15.42)$$

can be transformed into a Sturm–Liouville equation. Substituting  $a(x) = 1$  and  $b(x) = -2x$  into (15.41) yields

$$\xi(x) = \exp \left[ \int^x (-2s) ds \right] = e^{-x^2},$$

by which we multiplying both sides of (15.42), to obtain

$$e^{-x^2} y'' - 2x e^{-x^2} y' + 2\alpha e^{-x^2} y = \left( e^{-x^2} y' \right)' + 2\alpha e^{-x^2} y = 0,$$

This is the Sturm–Liouville form with

$$p(x) = e^{-x^2}, \quad q(x) = 0, \quad w(x) = e^{-x^2},$$

and  $\lambda = 2\alpha$ .

### 15.3.3 Self-adjoint Operators

We know many facts about Sturm–Liouville problems. Below is an important concept regarding the nature of these problems.

#### ♠ Adjoint operator:

The adjoint of an operator  $L$ , denoted by  $L^\dagger$ , is defined by

$$\int_a^b f^*(x) [Lg(x)] \rho(x) dx = \left\{ \int_a^b g^*(x) [L^\dagger f(x)] \rho(x) dx \right\}^*. \quad (15.43)$$

Using **inner product notation**, we can write the definition of the adjoint operator (15.43) as

$$(f, Lg) = (g, L^\dagger f).$$

The most important terminology in this section is given below.

♠ **Self-adjoint operator:**

An operator  $L$  is called **self-adjoint** (or **Hermitian**) if

$$L = L^\dagger$$

or, in inner product notation,

$$(f, Lg) = (g, Lf)^*.$$

It should be noted that an operator is said to be self-adjoint only if certain boundary conditions are met by the functions  $f$  and  $g$  on which it acts. An illustrative example follows:

*Examples* Let us derive the required boundary conditions for the linear operator

$$L = \frac{d^2}{dx^2}$$

to be self-adjoint over the interval  $[a, b]$ . From the definition of self-adjoint operators, the operator  $L$  should satisfy the relation:

$$\int_a^b f^* \frac{d^2 g}{dx^2} dx = \left( \int_a^b g^* \frac{d^2 f}{dx^2} dx \right)^*. \quad (15.44)$$

Through integration by parts, the left-hand side gives

$$\int_a^b f^* \frac{d^2 g}{dx^2} dx = \left[ f^* \frac{dg}{dx} \right]_a^b + \left[ -g \frac{df^*}{dx} \right]_a^b + \int_a^b g \frac{d^2 f^*}{dx^2} dx. \quad (15.45)$$

From a comparison as (15.44) and (15.45), it follows that the operator  $L$  is Hermitian provided that

$$\left[ f^* \frac{dg}{dx} \right]_a^b = \left[ g \frac{df^*}{dx} \right]_a^b.$$

### 15.3.4 Required Boundary Condition

In the example in Sect. 15.3.3, we derived the required boundary condition for a specific Sturm–Liouville operator to be self-adjoint. For general Sturm–Liouville operators, such a required boundary condition is given by the following theorem.

**♠ Theorem:**

A Sturm–Liouville operator is self-adjoint on  $[a, b]$  if any two eigenfunctions  $y_i$  and  $y_j$  of (15.38) satisfy the boundary condition

$$[py_i^* y_j']_a^b = 0. \quad (15.46)$$

*Proof* It follows from the explicit form of the Sturm–Liouville operator  $L$  that

$$(y_i, Ly_j) = -\frac{1}{w} \int_a^b y_i^* (py_j')' dx - \frac{1}{w} \int_a^b y_i^* q y_j dx. \quad (15.47)$$

The first integral is integrated by parts to give

$$-\frac{1}{w} [y_i^* p y_j']_a^b + \frac{1}{w} \int_a^b (y_i^*)' p y_j' dx,$$

in which the first term vanishes because we have assumed the boundary condition (15.46). Integration by parts then yields

$$\frac{1}{w} [(y_i^*)' p y_j]_a^b - \frac{1}{w} \int_a^b [(y_i^*)' p]' y_j dx,$$

where the first term is again zero owing to our assumption. As a result, the sum of integrals  $I$  in (15.47) reads

$$(y_i, Ly_j) = \frac{1}{w} \int_a^b \left\{ [-(y_i^*)' p]' y_j - y_i^* q y_j \right\} dx \quad (15.48)$$

$$= \frac{1}{w} \left\{ - \int_a^b [y_j^* (py_i')' - y_i q y_j^*] dx \right\}^* = (y_j, Ly_i)^*, \quad (15.49)$$

which completes the proof. ♣

### 15.3.5 Reality of Eigenvalues

**♠ Theorem:** For a Sturm–Liouville system under the boundary condition (15.46), we have:

- (a) All eigenvalues are real.
- (b) Eigenfunctions corresponding to distinct eigenvalues are orthogonal.

*Proof of (a).* If an eigenfunction  $y_n$  belongs to the eigenvalue  $\lambda_n$ , then

$$\begin{aligned}\lambda_n^*(y_n, y_n) &= (\lambda_n y_n, y_n) = -(Ly_n, y_n) \\ &= -(y_n, Ly_n) = \lambda_n(y_n, y_n).\end{aligned}$$

This indicates that  $\lambda_n^* = \lambda_n$  since  $(y_n, y_n) > 0$ . Therefore  $\lambda_n$  is real for all  $n$ . ♣

*Proof of (b).* According to the same argument as above,

$$\begin{aligned}\lambda_m(y_m, y_n) &= (\lambda_m y_m, y_n) = -(Ly_m, y_n) \\ &= -(y_m, Ly_n) = \lambda_n(y_m, y_n).\end{aligned}$$

Thus, for  $\lambda_m \neq \lambda_n$ ,  $(y_m, y_n) = 0$ , which means that eigenfunctions corresponding to distinct eigenvalues are orthogonal. ♣

**Remark.** If eigenvalues are degenerate, say,  $\lambda_m = \lambda_n$  ( $m \neq n$ ), an orthogonal set of eigenfunctions is constructed using the **Gram–Schmidt orthogonalization method**. Namely, we can choose the eigenfunctions to be orthogonal to each other with respect to the weight function  $w$  such that if  $(y_m, y_n) \neq 0$ , we replace  $y_n$  by  $\tilde{y}_n = y_n - ay_m$  where  $a$  should be chosen to be  $(y_m, \tilde{y}_n) = 0$ .

## Exercises

1. Show that the **Bessel equation** given by

$$x^2 y'' + xy' + (x^2 - n^2)y = 0 \quad \text{with } n \geq 0 \text{ and } x \in (-\infty, \infty)$$

can be expressed in the form of a Sturm–Liouville equation.

**Solution:** After the transformation  $x \rightarrow kx$ , we have

$$[xy'(kx)]' + \left(-\frac{n^2}{x} + k^2 x\right)y(kx) = 0, \quad n \geq 0,$$

where  $p = x$ ,  $q = -n^2x$ ,  $w = x$ , and the parameter  $\lambda = k^2$  in (15.38). ♣

2. The **Bernoulli equation** is given as a nonlinear equation by

$$y' = a(x)y + b(x)y^k, \tag{15.50}$$

where  $a(x)$ ,  $b(x)$  are continuous functions in an interval  $I$  and  $k$  is an arbitrary constant.

- (a) Show that the transformation  $u = y^{1-k}$  provides an inhomogeneous linear equation for  $u$ .

- (b) Find a solution for the transformed linear equation for  $u$  under the initial condition  $u(x_0) = u_0$ .

**Solution:**

- (a) The transformed equation becomes  $u' = (1-k)a(x)u + (1-k)b(x)$ .  
 (b) The above equation can be reduced to an inhomogeneous linear equation of the form

$$u' = p(x)u + q(x),$$

where  $p(x) = (1-k)a(x)$ ,  $q(x) = (1-k)b(x)$  are continuous functions. Let  $P(x)$  be a function whose derivative is  $p(x)$  such that

$$P(x) = \int_{x_0}^x p(t)dt,$$

where  $x_0$  is a fixed point in  $I$ . Multiplying both sides of (15.50) by  $e^{P(x)}$  to, we have the relation

$$(e^P u)' = e^P (u' - pu) = e^P q.$$

Therefore, we obtain a solution such that

$$u(x) = u_0 e^{-P(x)} + e^{-P(x)} \int_{x_0}^x e^{P(s)} q(s) ds,$$

where  $u_0$  comes from the initial condition. ♣

3. The **logistic equation** is a special type of Bernoulli equation given by

$$y' = ay - by^2, \quad (15.51)$$

where  $a, b$  are constants. Find a solution for the above by imposing the initial condition  $y(x_0) = y_0$ .

**Solution:** Using a solution for Exercise 2(b) by setting  $k = 2$ , we have

$$y(x) = \frac{a}{b + (a/y_0 - b)e^{-a(x-x_0)}}.$$

Note that  $y(x) = a/b$  as  $x \rightarrow \infty$ . ♣

4. The **Riccati equation** is a nonlinear equation given by

$$y' + p(x)y + q(x)y^2 = r(x). \quad (15.52)$$

- (a) Assuming  $u(x)$  to be a particular solution of the above, namely, a solution when we set  $r(x) = 0$ , show that  $z(x)$  defined by  $y(x) = u(x) + z(x)$  constitutes the Bernoulli equation.  
 (b) Show that the Riccati equation is reduced to a linear equation of the second order by the transformation  $y = Qv'/v$ .

**Solution:**

(a) Substituting  $y = u + z$  into the equation, we have

$$[z' + p(z^2 + 2uz) + qz] + [u' + pu + qu^2 - r] = 0.$$

The second parenthesis vanishes and we have the Bernoulli equation such that  $z' + (2up + q)z + pz^2 = 0$ .

(b) The first order derivative gives

$$y' = Q \left( \frac{v''}{v} - \frac{v'^2}{v^2} \right) + Q' \frac{v'}{v}.$$

Thus, we have

$$Q \frac{v''}{v} + (pQ - 1)Q \frac{v'^2}{v^2} + (Q' + qQ) \frac{v'}{v} + r = 0.$$

Setting  $Q = 1/p(x)$ , we have  $v'' + \left(q - \frac{p'}{p}\right)v' + prv = 0$ . ♣



## System of Ordinary Differential Equations

**Abstract** In this chapter we focus on an autonomous system (Sect. 16.3), which is a specific type of system of ordinary differential equations. Autonomous systems can be used to describe the dynamics of the physical objects that are encountered in physics and engineering problems, wherein the laws governing the motion of the objects are time-independent, namely, they hold true at all times. The stability of these dynamical systems is characterized by the critical point (Sect. 16.3.3), whose nature is revealed by the functional form of the autonomous systems.

### 16.1 Systems of ODEs

#### 16.1.1 Systems of the First-Order ODEs

This section deals with  $n$  coupled ordinary differential equations (ODEs). The formal definition is stated below.

♠ **Systems of ODEs:**

A system of ODEs is given by

$$\begin{aligned} & F_i \left[ x; y_1, y_1', y_1'', \dots, y_1^{(r_{i1})}; y_2, y_2', y_2'', \dots, y_2^{(r_{i2})}; \dots \right] \\ & = 0 \quad (i = 1, 2, \dots), \end{aligned} \tag{16.1}$$

which involves a set of unknown functions  $y_1(x), y_2(x), \dots$  and their derivatives with respect to a single independent variable  $x$ .

For each  $i$ th equation of (16.1), we denote the highest order of the derivatives of  $y_j$  by  $r_{ij}$ . Hereafter, we consider the case of  $r_{ij} \equiv 1$  for all  $i$  and  $j$ , i.e., a system of  $n$  ordinary differential equations (ODEs) of the first order expressed by

$$\begin{aligned}
y_1'(x) &= f_1(x, y_1, y_2, \dots, y_n), \\
y_2'(x) &= f_2(x, y_1, y_2, \dots, y_n), \\
&\dots \\
y_n'(x) &= f_n(x, y_1, y_2, \dots, y_n).
\end{aligned} \tag{16.2}$$

Here,  $\{f_k\}$ ,  $k = 1, 2, \dots, n$  are single-valued continuous functions in a certain domain of their arguments and  $\{y_k\}$ ,  $k = 1, 2, \dots, n$  are unknown complex functions of a real variable  $x$ .

### 16.1.2 Column-Vector Notation

For convenience we use **column-vector notation** for an ordered set of unknown functions  $\{y_k(x)\}$  in which each  $y_k(x)$  is called a component, which we denote by a bold-face letter:

$$\mathbf{y}(x) = [y_1(x), y_2(x), \dots, y_n(x)]^T, \tag{16.3}$$

where the norm of the vector is defined by

$$\|\mathbf{y}(x)\| = (|y_1|^2 + |y_2|^2 + \dots + |y_n|^2)^{1/2}. \tag{16.4}$$

Using vector notation, we can express (16.2) in the concise form

$$\mathbf{y}'(x) = \mathbf{f}(x, \mathbf{y}(x)), \tag{16.5}$$

where the column vector  $\mathbf{f}$  is defined by its components

$$\mathbf{f}(x, \mathbf{y}(x)) = [f_1, f_2, \dots, f_n]^T. \tag{16.6}$$

If there exists a set of functions  $\boldsymbol{\varphi}(x) = (\varphi_1(x), \varphi_2(x), \dots, \varphi_n(x))$  satisfying

$$\varphi_i(x)' = f_i(x, \varphi_1(x), \varphi_2(x), \dots, \varphi_n(x)), \quad i = 1, 2, \dots, n,$$

we say  $\boldsymbol{\varphi}(x)$  is a solution of (16.2). The initial value problem consists of finding a solution  $\boldsymbol{\varphi}(x)$  of (16.5) in  $I$  satisfying the initial condition  $\boldsymbol{\varphi}(x_0) = \mathbf{y}_0 = (y_{10}, y_{20}, \dots, y_{n0})$ .

### 16.1.3 Reducing the Order of ODEs

Let consider an  $n$ th order ODE of  $u(x)$  given by

$$\frac{d^n u(x)}{dx^n} + p_1(x) \frac{d^{n-1} u(x)}{dx^{n-1}} + \dots + p_n(x) u(x) = q(x). \tag{16.7}$$

We show that equation (16.7) can always be reduced to a system of  $n$  first-order differential equations, which is stated as follows:

♠ **Theorem:**

Given an  $n$ th-order ODE, it can always be reduced to a system of  $n$  first-order ODEs.

*Proof* We take  $u(x)$  and its derivatives  $u', u'', \dots, u^{(n-1)}$  as new unknown functions defined by

$$y_k(x) \equiv \frac{d^{k-1}u(x)}{dx^{k-1}}, \quad k = 1, 2, \dots, n. \quad (16.8)$$

It is evident that (16.7) is equivalent to the following set of equations:

$$y'_1 = y_2, \quad y'_2 = y_3, \quad \dots, \quad y'_{n-1} = y_n \quad (16.9)$$

and

$$y'_n = -p_1 y_n - p_2 y_{n-1} - \dots - p_n y_1 + q. \quad (16.10)$$

Equations (16.9) and (16.10) can be written in a brief vector form as

$$\frac{d\mathbf{y}(x)}{dx} = \mathbf{f}(x, \mathbf{y}), \quad (16.11)$$

where the column vectors are defined as

$$\mathbf{y} = (y_1, y_2, \dots, y_n)$$

and

$$\begin{aligned} \mathbf{f} &= \mathbf{f}(x, \mathbf{y}) \\ &= [y_2, y_3, \dots, y_n, -p_1 y_n - p_2 y_{n-1} - \dots - p_n y_1 + q]^T. \quad \clubsuit \end{aligned}$$

*Example* One of the most famous systems of the type (16.11) results from the equation of motion for a particle of mass  $m$ . For a mobile particle along the  $x$ -axis, the equation of motion is

$$m \frac{d^2 x(t)}{dt^2} = F\left(t, x(t), \frac{dx(t)}{dt}\right), \quad (16.12)$$

where  $t$  is the time and  $F$  represents the force acting on the particle. To see how the second-order ODE (16.12) can be viewed as a system of the form (16.11), we make the following substitutions:

$$t \rightarrow x, \quad x \rightarrow y_1, \quad \frac{dx}{dt} \rightarrow y_2.$$

Then (16.12) is equivalent to a system of two equations:

$$\begin{aligned} y'_1 &= y_2, \\ y'_2 &= \frac{1}{m} F(x, y_1, y_2), \end{aligned}$$

which is of the form of  $\mathbf{y}'(x) = \mathbf{f}(x, \mathbf{y})$ .

### 16.1.4 Lipschitz Condition in Vector Spaces

The vector equation

$$\mathbf{y}'(x) = \mathbf{f}(x, \mathbf{y}) \quad (16.13)$$

is obviously analogous to the scalar equation

$$y'(x) = f(x, y).$$

This analogy implies the possibility that the definition of a Lipschitz condition can be extended to the vector equation. The extended Lipschitz condition provides a simple sufficient condition for the uniqueness and existence of solutions, which implies that all the theorems for the scalar equation can be generalized so as to hold for the vector equation.

♠ **Lipschitz condition for a vector function:**

A vector function  $\mathbf{f}(x, \mathbf{y})$  in (16.13) is said to satisfy the Lipschitz condition on a region  $R$  if and only if

$$\begin{aligned} |\mathbf{f}(x, \mathbf{y}(x)) - \mathbf{f}(x, \mathbf{z}(x))| &\leq K |\mathbf{y}(x) - \mathbf{z}(x)|, \\ (R : |x - x_0| \leq a, \quad |\mathbf{y} - \mathbf{y}_0| \leq b, \quad |\mathbf{z} - \mathbf{z}_0| \leq b). \end{aligned} \quad (16.14)$$

for the Lipschitz constant  $K$ .

When  $\mathbf{f}(x, \mathbf{y})$  satisfies the Lipschitz condition noted above, we see from (16.14) that

$$\begin{aligned} &|f_k(x, y_1, y_2, \dots, y_n) - f_k(x, z_1, z_2, \dots, z_n)| \\ &\leq K(|y_1 - z_1| + |y_2 - z_2| + \dots + |y_n - z_n|) \quad (k = 1, 2, \dots, n). \end{aligned} \quad (16.15)$$

Using this, we can prove the theorem of the existence and uniqueness of solutions for the general vector equation (16.13). For instance, the uniqueness of the solution for (16.13) is straightforward as shown below. The right-hand side of (16.15) yields

$$\begin{aligned} K \sum_{k=1}^n |y_k(x) - z_k(x)| &\leq K \sum_{k=1}^n \int_{x_0}^x |f_k(x, \mathbf{y}(x)) - f_k(x, \mathbf{z}(x))| dx \\ &\leq nK^2 \int_{x_0}^x \sum_{k=1}^n |y_k(x) - z_k(x)| dx, \end{aligned} \quad (16.16)$$

which holds for the interval  $I$ ;  $x \in [x_0, x_0 + \delta]$  for any small  $\delta$ . Since the left-hand side of (16.16) is continuous on  $I$ , it has a maximum at some  $x$ , which we label  $\mu$ . Then, the inequality (16.16) becomes

$$\mu \leq nK\mu(x - x_0) \leq nK\mu\delta,$$

which gives us  $\mu(1 - nK\delta) \leq 0$ . For any small  $\delta > 0$ , we have  $\mu = 0$ , which indicates that  $\sum |y_k - z_k| = 0$ . The same holds true for the case  $x < x_0$ . Thus, the solution of (16.13) is unique.

## Exercises

1. Consider a initial value problem given by

$$\mathbf{y}' = \mathbf{f}(x, \mathbf{y}), \quad \mathbf{y}(x_0) = \mathbf{y}_0,$$

defined on  $\mathbf{R} : |x - x_0| \leq a, |\mathbf{y} - \mathbf{y}_0| \leq b, (a, b > 0)$ . Assuming that  $\mathbf{f}$  is continuous on  $\mathbf{R}$ , a sequence of successive approximations  $\varphi_0, \varphi_1, \dots$  is given by

$$\varphi_0(x) = \mathbf{y}_0$$

and

$$\varphi_{n+1}(x) = \mathbf{y}_0 + \int_{x_0}^x \mathbf{f}(t, \varphi_n(t)) dt \quad \text{for } n = 1, 2, \dots.$$

Using this procedure, find a sequence of successive approximations for

$$(y'_1, y'_2) = (y_2, -y_1), \quad \text{for } \mathbf{y}(0) = (0, 1).$$

**Solution:** Here  $\mathbf{f}(x, \mathbf{y}) = (y_2, -y_1)$ , so we have

$$\varphi_0(x) = (0, 1),$$

$$\varphi_1(x) = (0, 1) + \int_0^x (1, 0) dt = (x, 1),$$

$$\varphi_2(x) = (0, 1) + \int_0^x (1, -t) dt = (0, 1) + \left(x, -\frac{x^2}{2}\right) = \left(x, 1 - \frac{x^2}{2}\right).$$

Continuing with this process, we find the solution of the problem as  $\varphi_k(x) \rightarrow \varphi(x) = (\sin x, \cos x)$ . ♣

## 16.2 Linear System of ODEs

### 16.2.1 Basic Terminology

We now focus on a particular class of systems of ODEs called a **linear** system of first-order ODEs, described by

$$\frac{dy_1(x)}{dx} - \sum_{j=1}^n a_{1j}(x)y_j(x) = q_1(x),$$

$$\begin{aligned} \frac{dy_2(x)}{dx} - \sum_{j=1}^n a_{2j}(x)y_j(x) &= q_2(x), \\ \dots \\ \frac{dy_n(x)}{dx} - \sum_{j=1}^n a_{nj}(x)y_j(x) &= q_n(x). \end{aligned}$$

Here  $a_{kj}(x)$  and  $q_k(x)$  with  $j, k = 1, 2, \dots, n$  are continuous functions on  $x$  on some interval  $I$ . For convenience we use the vector representation given by

$$\frac{d\mathbf{y}(x)}{dx} - \mathbf{A}(x)\mathbf{y}(x) = \mathbf{q}(x), \quad (16.17)$$

where  $\mathbf{A} = [a_{kj}]$  is an  $n \times n$  matrix. Therefore,  $\mathbf{A}\mathbf{y}$  stands for the matrix  $\mathbf{A}$  applied to the column vector  $\mathbf{y} = [y_1, y_2, \dots, y_n]^T$ , namely, the **linear transform** of  $\mathbf{y}$  by  $\mathbf{A}$ . The vector  $\mathbf{q}$  is defined as  $\mathbf{q} = [q_1, q_2, \dots, q_n]^T$ . Given any  $\mathbf{y}(x_0)$  for  $x_0$  in  $I$ , there exists a unique solution  $\boldsymbol{\varphi}(x)$  on  $I$  such that  $\boldsymbol{\varphi}(x_0) = [y_1(x_0), y_2(x_0), \dots, y_n(x_0)]^T$ .

The use of the linear operator  $L$  to (16.17) yields

$$L[\mathbf{y}(x)] = \mathbf{q}(x),$$

where the  $L$  is defined as

$$L = \frac{d}{dx} - \mathbf{A}. \quad (16.18)$$

If  $\mathbf{q}(x) = \mathbf{0}$  for all  $x$  on  $I$ , (16.17) is said to be a **linear homogeneous system** of  $n$ th order, expressed by

$$\frac{d\mathbf{y}(x)}{dx} - \mathbf{A}(x)\mathbf{y}(x) = \mathbf{0}. \quad (16.19)$$

Otherwise, (16.17) is called **inhomogeneous**. A homogeneous system obtained from the inhomogeneous system (16.17) by setting  $\mathbf{q}(x) \equiv \mathbf{0}$  is called the **reduced** or **complementary system**.

**Remark.** Note that every linear homogeneous system always has a trivial solution  $\boldsymbol{\varphi}(x) = \mathbf{0}$ , as can be immediately checked. From the uniqueness of the solution, therefore, there is no solution vanishing at *only* some point of  $x$ .

### 16.2.2 Vector Space of Solutions

Let  $\boldsymbol{\varphi}_i(x)$  ( $i = 1, 2, \dots$ ) be solutions for an  $n$ -dimensional linear homogeneous system

$$\mathbf{y}'(x) = \mathbf{A}(x)\mathbf{y}(x). \quad (16.20)$$

Referring to the axioms given in Sect. 4.2.1, it readily follows that the solutions  $\{\boldsymbol{\varphi}_i(x)\}$  form a vector space  $V$ . Indeed, if  $\boldsymbol{\varphi}_1(x)$  and  $\boldsymbol{\varphi}_2(x)$  are solutions of (16.20), then  $c_1\boldsymbol{\varphi}_1(x) + c_2\boldsymbol{\varphi}_2(x)$  with arbitrary constants  $c_1, c_2$  is also a solution of (16.20), and so on.

Now we pose a question as to the dimension of the vector space  $V$  mentioned above. We have the answer in the following theorem:

♠ **Theorem:**

Solutions of the system (16.20) on an interval  $I$  form an  $n$ -dimensional vector space if the  $n \times n$  matrix  $\mathbf{A}(x)$  is continuous on  $I$ .

*Proof* The continuity of  $\mathbf{A}(x)$  implies that all its components do not diverge. This allows us to set a constant  $K$ ,

$$K = \max \sum_{i=1}^n |a_{ij}(x)|,$$

and it then follows that the vector  $\mathbf{f}$  defined by  $\mathbf{f}(x) = \mathbf{A}(x)\mathbf{y}(x)$  satisfies the Lipschitz condition:

$$|\mathbf{f}(x, \mathbf{y}) - \mathbf{f}(x, \mathbf{z})| \leq K |\mathbf{y} - \mathbf{z}| \quad \text{for } x \in I.$$

From the existence and uniqueness theorems we know that there are  $n$  solutions  $\varphi_i(x)$  of (16.20) such that each solution exists on the entire interval  $I$  and satisfies the initial condition

$$\varphi_i(x_0) = \mathbf{e}_i \quad (i = 1, 2, \dots, n) \quad \text{for } x_0 \in I, \quad (16.21)$$

where the  $\mathbf{e}_i$ 's are  $n$  linearly independent vectors.

We tentatively assume that the solutions  $\varphi_i$  are linearly *dependent* on  $I$ . Then there exist constants  $c_i$ , not all zero, such that

$$\sum_{i=1}^n c_i \varphi_i(x) = 0 \quad \text{for every } x \text{ on } I.$$

In particular, setting  $x = x_0$ , and using the initial condition (16.21), we have

$$\sum_{i=1}^n c_i \mathbf{e}_i = 0,$$

which contradicts the assumed linear independence of  $\mathbf{e}_i$ . Hence, we conclude that the solutions  $\varphi_i$  are linearly *independent* on  $I$ .

Next we prove the completeness of  $\{\varphi_i(x)\}$ ; i.e., that every solution  $\psi(x)$  of (16.20) can be expanded as a linear combination of  $\varphi_i(x)$  satisfying the initial condition (16.21). Since the  $\mathbf{e}_i$  are linearly independent in the  $n$ -dimensional Euclidean space  $E_n$ , they form a basis for  $E_n$  and there exist unique constants  $b_i$  such that the constant vector  $\psi(x_0)$  can be expressed as

$$\psi(x_0) = \sum_{i=1}^n b_i \mathbf{e}_i. \quad (16.22)$$

Consider the vector

$$\varphi(x) = \sum_{i=1}^n b_i \varphi_i(x),$$

where the  $b_i$  are identical to those in (16.22). Clearly  $\varphi(x)$  is a solution of (16.20) on  $I$ . In addition, the initial value of  $\varphi$  reads

$$\varphi(x_0) = \sum_{i=1}^n b_i e_i,$$

so that  $\varphi(x_0) = \psi(x_0)$ . In view of the uniqueness theorem, we have

$$\varphi(x) = \psi(x) \quad \text{for every } x \text{ on } I.$$

This leads to the conclusion that every solution  $\psi(x)$  of an  $n$ th-order linear homogeneous system (16.20) is expressed by the unique linear combination

$$\psi(x) = \sum_{i=1}^n b_i \varphi_i(x) \quad \text{for every } x \text{ on } I,$$

where the  $b_i$  are uniquely determined once we have  $\psi(x)$ . As a result,  $n$  solutions  $\varphi_i(x)$  of the system (16.20) form the basis for an  $n$ -dimensional vector space. ♣

### 16.2.3 Fundamental Systems of Solutions

Again let  $\varphi_i(x) = [\varphi_{1i}(x), \dots, \varphi_{ni}(x)]^T$  ( $i = 1, 2, \dots, n$ ) be solutions of the linear homogeneous system (16.20) such that

$$\varphi_i(x)' = \mathbf{A}(x)\varphi_i(x) \quad \text{for all } i = 1, 2, \dots, n.$$

Note here that  $\{\varphi_i(x)\}$  may or may not be linearly independent, since no initial condition is imposed (contrary to the case of (16.21)). Specifically, if the set  $\{\varphi_i(x)\}$  is endowed with the linear independence property, it is called the **fundamental system of solutions** of (16.20).

#### ♠ Fundamental system of solutions:

A collection of  $n$  solutions  $\{\varphi_i(x)\}$  of an  $n$ -dimensional linear homogeneous system is called a **fundamental system of solutions** of the system if it is linearly independent.

**Remark.** The significance of a fundamental system of solutions lies in the fact that it can describe any solution  $\varphi(x)$  of the corresponding linear homogeneous system. Consequently, the problem of finding a solution  $\varphi(x)$  becomes equivalent to that of finding  $n$  linearly independent solutions.

With this terminology, the theorem presented in Sect. 16.2.2 leads to the following result:

♠ **Theorem:**

A fundamental system of solutions exists for an arbitrary linear homogeneous system.

*Example* The second-order equation

$$y''(t) + y(t) = 0 \quad (16.23)$$

is equivalent to the two-dimensional linear system

$$\mathbf{u}'(t) = \mathbf{A}\mathbf{u}(t) \quad (16.24)$$

with

$$\mathbf{u}(t) = \begin{bmatrix} y'(t) \\ y(t) \end{bmatrix} \quad \text{and} \quad \mathbf{A} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

The fundamental system of solutions of (16.24) is given by

$$\boldsymbol{\varphi}_1(t) = [\cos t, -\sin t]^T \quad \text{and} \quad \boldsymbol{\varphi}_2(t) = [\sin t, \cos t]^T,$$

whose linear independence follows from the fact that  $c_1 \sin t \pm c_2 \cos t \equiv 0$  implies  $c_1 = c_2 = 0$ . Furthermore,  $\boldsymbol{\varphi}_1(0) = (1, 0)$  and  $\boldsymbol{\varphi}_2(0) = (0, 1)$ , so any solution  $\boldsymbol{\varphi}(t)$  is given by

$$\boldsymbol{\varphi}(t) = a_0 \boldsymbol{\varphi}_1(t) + b_0 \boldsymbol{\varphi}_2(t) \quad \text{for } -\infty < t < \infty, \quad (16.25)$$

where  $\boldsymbol{\varphi}(0) = (a_0, b_0)$ .

**Remark.** The solution  $\boldsymbol{\varphi}(t)$  in (16.25) corresponds to the solution of the second-order ODE (16.23) satisfying the initial conditions:  $y(0) = a_0$  and  $y'(0) = b_0$ .

#### 16.2.4 Wronskian for a System of ODEs

The theorems given in Sect. 16.2.2 and 16.2.3 ensure the existence of a fundamental system of solutions for any linear homogeneous system of the form

$$\mathbf{y}'(x) = \mathbf{A}(x)\mathbf{y}(x). \quad (16.26)$$

However, it provides no information as to whether a certain set of solutions is a fundamental system or not. In what follows, we consider the criteria concerning this issue. Following are preliminary concepts that we need in order to proceed.

♠ **Wronsky determinant:**

Let  $\{\varphi_k(x)\}$  ( $k = 1, 2, \dots, n$ ) be solutions of (16.26), where  $\varphi_k(x) = [\varphi_{1k}(x), \dots, \varphi_{nk}(x)]^T$ . Then the scalar function

$$W(x) = \det \begin{bmatrix} \varphi_{11} & \varphi_{12} & \cdots & \varphi_{1n} \\ \varphi_{21} & \varphi_{22} & \cdots & \varphi_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \varphi_{n1} & \varphi_{n2} & \cdots & \varphi_{nn} \end{bmatrix} \quad (16.27)$$

is called the **Wronsky determinant** (or the **Wronskian**) of the solutions  $\{\varphi_k(x)\}$ .

If  $\{\varphi_k(x)\}$  is a fundamental system of solutions of (16.26), then the matrix corresponding to  $W(x)$  is called a **fundamental matrix**. Hence, a fundamental matrix is a matrix whose columns form a fundamental system of solutions of (16.26).

*Example* For the two-dimensional system given in Sect. 16.2.3, the matrix

$$\Phi(t) = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix}, \quad -\infty < t < \infty$$

is a fundamental matrix and  $W(t) \equiv 1$  for all  $t$ .

### 16.2.5 Liouville Formula for a Wronskian

The following theorem shows that given any  $n$  solutions of (16.26) and any  $t_0$  in  $(r_1, r_2)$ , we can completely determine the corresponding Wronskian without computing the  $n \times n$  determinant.

♠ **Liouville formula:**

Let  $\{\varphi_k(x)\}$  ( $k = 1, 2, \dots, n$ ) be any  $n$  solution of (16.26) and let  $x_0$  be in  $(r_1, r_2)$ . Then the Wronskian of  $\{\varphi_k(x)\}$  for  $x \in (r_1, r_2)$  is given by

$$W(x) = W(x_0) \exp \left[ \int_{x_0}^x \operatorname{tr} \mathbf{A}(s) ds \right].$$

See Exercise 2 for the proof. Since  $\exp \left[ \int_{x_0}^x \operatorname{tr} \mathbf{A}(s) ds \right]$  is never zero, the theorem implies that the Wronskian of any collection of  $n$  solutions of (16.26)

is *identically zero* or *never zero* on  $(r_1, r_2)$ . The latter case characterizes a fundamental system, as shown by the following theorem:

♠ **Theorem:**

A necessary and sufficient condition for  $\{\varphi_k(x)\}$  ( $k = 1, 2, \dots, n$ ) to be a fundamental system of solutions of (16.26) is that  $W(x) \neq 0$  for  $r_1 < x < r_2$ .

*Proof* Let  $\{\varphi_k(x)\}$  ( $k = 1, 2, \dots, n$ ) be a fundamental system of solutions of (16.26) and let  $\varphi(x)$  be any nontrivial solution. Then there exist  $c_1, \dots, c_n$  not all zero such that  $\varphi(x) = \sum_{i=1}^n c_i \varphi_i(x)$ , and by the uniqueness of the solutions the  $c_i$  are unique. If  $\mathbf{c} = [c_1, \dots, c_n]^T$  and  $\Phi(x)$  is the fundamental matrix of  $\{\varphi_k(x)\}$ , then the previous relation can be written as

$$\varphi(x) = \mathbf{c}\Phi(x).$$

For any  $x$  in  $(r_1, r_2)$ , this is a system of  $n$  linear equations in the unknowns  $c_1, \dots, c_n$ . Since this has a unique solution in  $\mathbf{c}$ ,  $\det \Phi$  cannot be zero, i.e.,

$$\det \Phi(x) = W(x) \neq 0 \quad \text{for any } x \in (r_1, r_2).$$

Conversely,  $W(x) \neq 0$  for  $r_1 < x < r_2$ , implies that the columns  $\varphi_1(x), \dots, \varphi_n(x)$  of  $\Phi(x)$  are linearly independent for  $r_1 < x < r_2$ . Since they are solutions of (16.26), they form a fundamental system of solutions. ♣

### 16.2.6 Wronskian for an $n$ th-Order Linear ODE

The previous results for systems of ODEs can be applied to an  $n$ th-order linear equation

$$u^{(n)}(x) + a_1(x)u^{(n-1)}(x) + \dots + a_n(x)u(x) = 0, \quad (16.28)$$

since (16.28) is transformed into a vector form as

$$\mathbf{y}' = \mathbf{A}\mathbf{y}, \quad (16.29)$$

where

$$\mathbf{y} = \begin{bmatrix} u \\ u' \\ \cdots \\ \cdots \\ u^{(n-1)} \end{bmatrix} \quad \text{and} \quad \mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdots & & & & \cdots \\ 0 & & & 0 & 1 \\ -a_n(x) & -a_{n-1}(x) & \cdots & -a_2(x) & -a_1(x) \end{bmatrix}.$$

Relevant terminology and theorems are given below.

♠ **Fundamental system of solutions:**

A collection

$$\xi_1(x), \dots, \xi_n(x), \quad r_1 < x < r_2$$

of solutions of (16.28) is called a **fundamental system of solutions** of (16.28) if it is linearly independent.

♠ **Theorem:**

A fundamental system of solutions of equation (16.28) exists.

*Proof* We know that a fundamental system of solutions of (16.29) exists, and we express it by  $\varphi_1(x), \dots, \varphi_n(x)$ , where  $\varphi_k(x) = [\varphi_{1k}(x), \dots, \varphi_{nk}(x)]^T$ . Furthermore, we may assume that given  $x_0$  in  $(r_1, r_2)$ ,

$$\varphi_k(x_0) = [0, \dots, 0, 1, 0, \dots, 0]^T \equiv \mathbf{e}_k, \quad k = 1, 2, \dots, n,$$

where the single nonzero component 1 in  $\mathbf{e}_k$  is assigned to the  $k$ th place in the square brackets. By the correspondence of solutions of (16.28) and (16.29), we have

$$\varphi_k(x) = [\xi_k(x), \xi'_k(x), \dots, \xi_k^{(n-1)}(x)]^T$$

for some solution  $u(x) = \xi_k(x)$  of (16.28). The collection  $\xi_1(x), \dots, \xi_n(x)$  comprises distinct nontrivial solutions, since they satisfy distinct initial conditions and  $\xi_k \equiv 0$  for  $r_1 < x < r_2$  would imply that  $\varphi_k(x) \equiv \mathbf{0}$ , which is impossible.

Finally, if there existed constants  $c_1, \dots, c_n$  not all zero such that  $\sum_{k=1}^n c_k \xi_k(x) \equiv 0$  for  $r_1 < x < r_2$ , then

$$\sum_{k=1}^n c_k \xi'_k(x) \equiv 0, \dots, \sum_{k=1}^n c_k \xi_k^{(n-1)}(x) \equiv 0, \quad r_1 < x < r_2.$$

This implies that

$$\sum_{k=1}^n c_k \varphi_k(x) \equiv \mathbf{0}, \quad r_1 < x < r_2,$$

which contradicts the fact that  $\{\varphi_k(x)\}$  is a fundamental system of (16.29).

♣

We now define the Wronskian of a collection of  $n$  solutions of (16.28).

♠ **Wronsky determinant:**

Given any collection  $\xi_1(x), \dots, \xi_n(x)$  of solutions of (16.28), then

$$W(x) = \det \begin{bmatrix} \xi_1 & \xi_2 & \cdots & \xi_n \\ \xi_1' & \xi_2' & \cdots & \xi_n' \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \xi_1^{(n-1)} & \xi_2^{(n-1)} & \cdots & \xi_n^{(n-1)} \end{bmatrix} \quad (16.30)$$

is called the **Wronsky determinant** (or the **Wronskian**) of the solutions  $\{\xi_k(x)\}$  ( $k = 1, 2, \dots, n$ ).

As before, if  $\xi_1(x), \dots, \xi_n(x)$  make up a fundamental system of (16.28), then the matrix corresponding to  $W(x)$  is called a **fundamental matrix**. In any case, note that the columns of the matrix corresponding to  $W(x)$  are  $n$  solutions of the system (16.29). We may therefore immediately state a result analogous to the Liouville formula given in Sect. 16.2.4, noting that  $\text{tr} \mathbf{A}(x) = -a_1(x)$ :

♠ **Theorem:**

The Wronskian  $W(x)$  of any collection  $\xi_1(x), \dots, \xi_n(x)$  of solutions of (16.28) satisfies the relation

$$W(x) = W(x_0) \exp \left[ - \int_{x_0}^x a_1(s) ds \right], \quad r_1 < x_0, \quad x < r_2.$$

Finally, we have the result corresponding to the theorem in Sect. 16.2.4, for which the proof is virtually the same.

♠ **Theorem:**

A necessary and sufficient condition for  $\xi_1(x), \dots, \xi_n(x)$  to be a fundamental system of solutions of equation (16.28) is that

$$W(x) \neq 0 \quad \text{for } r_1 < x < r_2.$$

*Example* Assume a second-order equation

$$y''(x) + a(x)y(x) = 0.$$

For any two solutions  $\xi_1(x)$  and  $\xi_2(x)$ , we have

$$W(x) = \det \begin{pmatrix} \xi_1(x) & \xi_2(x) \\ \xi_1'(x) & \xi_2'(x) \end{pmatrix} = \text{const.}$$

The constant is nonzero if and only if  $\xi_1$  and  $\xi_2$  are linearly independent.

*Remark.* The fact that linear independence implies a nonvanishing Wronskian is a property of solutions of *linear* equations; i.e., it does not hold for *nonlinear* equations. To see this, we consider the functions  $\xi_1(x) = x^3$  and  $\xi_2(x) = |x|^3$ . They are linearly independent on  $-\infty < x < \infty$ , but

$$W(x) = \det \begin{pmatrix} x^3 & |x|^3 \\ 3x^2 & 3x|x| \end{pmatrix} = 0.$$

This results from the fact that  $\xi_1(x)$  and  $\xi_2(x)$  cannot both be solutions near  $x = 0$  of a second-order linear equation. In fact, they both satisfy  $\xi(0) = \xi'(0) = 0$  yet are distinct, which violates uniqueness.

### 16.2.7 Particular Solution of an Inhomogeneous System

We close this section by discussing an inhomogeneous linear equation

$$\frac{d\mathbf{y}(x)}{dx} - \mathbf{A}(x)\mathbf{y}(x) = \mathbf{q}(x). \quad (16.31)$$

Let  $\mathbf{q}(x)$  be continuous on  $x$  on some interval  $I$  and let  $\{\varphi_k\}$  ( $k = 1, 2, \dots, n$ ) be a fundamental system of solutions for the reduced equation of (16.31). A general solution of (16.31) can be written as the sum

$$\psi(x) = \varphi_p(x) + c_1\varphi_1(x) + \dots + c_n\varphi_n(x), \quad (16.32)$$

where  $\varphi_p(x)$  is a **particular solution** of (16.31) with no adjustable parameter.

A particular solution can be obtained from a fundamental system  $\{\varphi_k\}$  ( $k = 1, 2, \dots, n$ ) of the reduced equation (16.19) by means of the **method of variation of constant parameters**. We assume a particular solution of the form

$$\varphi_p(x) = C_1(x)\varphi_1(x) + \dots + C_n(x)\varphi_n(x), \quad (16.33)$$

where the coefficients  $\{C_k(x)\}$  ( $k = 1, 2, \dots, n$ ) are not constants, but unknown functions of  $x$ . Differentiating (16.33) on  $x$  and substituting it into (16.31), we obtain

$$\sum_{k=1}^n [C_k(x)\varphi_k'(x) + C_k'(x)\varphi_k(x) - C_k(x)\mathbf{A}(x)\varphi_k(x)] = \mathbf{q}(x). \quad (16.34)$$

Since  $\{\varphi_k\}$  ( $k = 1, 2, \dots, n$ ) are solutions of the reduced equation (16.19), equation (16.34) yields

$$\sum_{k=1}^n \varphi_k(x) C'_k(x) = \mathbf{q}(x). \quad (16.35)$$

If we express  $\varphi_k(x)$  by its components as

$$\varphi_k(x) = [\varphi_{k1}(x), \varphi_{k2}(x), \dots, \varphi_{kn}(x)],$$

equation (16.35) becomes

$$\sum_{j=1}^n \varphi_{kj}(x) C'_j(x) = q_k(x), \quad (16.36)$$

or equivalently,

$$\begin{bmatrix} \varphi_{11} & \varphi_{12} & \cdots & \varphi_{1n} \\ \varphi_{21} & \varphi_{22} & \cdots & \varphi_{2n} \\ & & \cdots & \\ & & \cdots & \\ \varphi_{n1} & \varphi_{n2} & \cdots & \varphi_{nn} \end{bmatrix} \begin{bmatrix} C'_1(x) \\ C'_2(x) \\ \cdot \\ \cdot \\ C'_n(x) \end{bmatrix} = \begin{bmatrix} q'_1(x) \\ q'_2(x) \\ \cdot \\ \cdot \\ q'_n(x) \end{bmatrix}. \quad (16.37)$$

The matrix  $[\varphi_{kj}]$  on the left-hand side of (16.37) satisfies  $\det[\varphi_{kj}] \neq 0$  because of the linear independence of the fundamental system of solutions  $\{\varphi_k\}$ . Hence, multiplying the **inverse matrix** (see Sect. 18.1.7) of  $[\varphi_{kj}]$  by the both sides of (16.37), we have

$$C'_k(x) = p_k(x), \quad (16.38)$$

where  $\{p_k(x)\}$ ,  $k = 1, 2, \dots, n$  are continuous functions obtained from (16.37). Thus once the differential equation (16.38) is solved with respect to  $C_k(x)$ , the solutions determine a particular solution of the form

$$\varphi_p(x) = \sum_{k=1}^n C_k(x) \varphi_k(x).$$

## Exercises

1. Suppose  $\varphi_1(x), \varphi_2(x)$  to be two solutions of the ODE  $y'' + a_1 y' + a_2 y = 0$  on an interval  $I$  containing a point  $x_0$ . Show that

$$W(\varphi_1, \varphi_2)(x) = e^{-a_1(x-x_0)} W(\varphi_1, \varphi_2)(x_0).$$

**Solution:** We have  $\varphi_1'' + a_1 \varphi_1' + a_2 \varphi_1 = 0$  and  $\varphi_2'' + a_1 \varphi_2' + a_2 \varphi_2 = 0$ . Multiplying the first equation by  $-\varphi_2$ , and the second by  $\varphi_1$  and adding we obtain

$$(\varphi_1 \varphi_2'' - \varphi_1'' \varphi_2) + a_1 (\varphi_1 \varphi_2' - \varphi_1' \varphi_2) = 0.$$

Note that  $W = \varphi_1\varphi_2' - \varphi_1'\varphi_2$  and  $W' = \varphi_1\varphi_2'' - \varphi_1''\varphi_2$ . Thus  $W$  satisfies the first-order equation:

$$W' + a_1W = 0,$$

which implies  $W(x) = ce^{-a_1x}$  in which  $c$  is some constant. Setting  $x = x_0$ , we have  $c = e^{-a_1x_0}W(x_0)$ , and thus

$$W(x) = e^{-a_1(x-x_0)}W(x_0). \quad \clubsuit$$

2. Assume an  $n$ -dimensional linear homogeneous system  $\mathbf{y}'(x) = \mathbf{A}(x)\mathbf{y}(x)$  on  $I = (a, b)$ , and let  $\{\mathbf{g}_i(x)\}$  be any  $n$  solution. Show that the Wronskian of  $\{\mathbf{g}_i(x)\}$  is given by

$$W(x) = W(x_0) \exp \left[ \int_{x_0}^x \text{tr} \mathbf{A}(s) ds \right], \quad \text{where } a < x_0 < b, \quad (16.39)$$

which is called the **Liouville formula**.

**Solution:** We show that  $W(x)$  satisfies the differential equation  $W'(x) = \text{tr} \mathbf{A}(x)W(x)$  from which the conclusion (16.39) follows. The expansion by **cofactors** of  $W(x)$  yields

$$W(x) = \sum_{j=1}^n \varphi_{ij}(x) \Delta_{ij}(x), \quad (16.40)$$

where  $\varphi_{ij}(x)$  is the  $j$ th element of  $\boldsymbol{\varphi}_i(x)$  and  $\Delta_{ij}(x)$  is the cofactor of  $W(x)$  (see Sect. 18.1.7 for the definition of the cofactor). Note that  $\Delta_{ij}(x)$  does not contain the term  $\varphi_{ij}(x)$ . Hence, if  $W(x)$  given in (16.40) is regarded as a function of the  $\varphi_{ij}(x)$ , we have  $\partial W / \partial \varphi_{ij} = \Delta_{ij}(x)$  and, by the chain rule,

$$W(x)' = \sum_{i,j=1}^n \frac{\partial W}{\partial \varphi_{ij}} \varphi_{ij}(x)' = \sum_{i=1}^n \left[ \sum_{j=1}^n \varphi_{ij}(x)' \Delta_{ij}(x) \right]. \quad (16.41)$$

We define  $W_i(x)$  as

$$W_i(x) \equiv \det \begin{pmatrix} \varphi_{11}(x) & \cdots & \varphi_{1n}(x) \\ \vdots & & \vdots \\ \varphi_{i1}(x)' & \cdots & \varphi_{in}(x)' \\ \vdots & & \vdots \\ \varphi_{n1}(x) & \cdots & \varphi_{nn}(x) \end{pmatrix},$$

where all the elements in the  $i$ th row are differentiated. Then, the expression in the square brackets in (16.41) is the expansion of  $W_i(x)$  by cofactors, so that

$$W(x)' = \sum_{i=1}^n W_i(x). \quad (16.42)$$

Furthermore, since  $\varphi_{ij}(x)' = \sum_{k=1}^n a_{ik}(x)\varphi_{kj}(x)$ , we have

$$W_i(x) \equiv \det \begin{pmatrix} \varphi_{11}(x) & \cdots & \varphi_{1n}(x) \\ \vdots & & \vdots \\ \sum_{k=1}^n a_{ik}\varphi_{k1}(x) & \cdots & \sum_{k=1}^n a_{ik}\varphi_{kn}(x) \\ \vdots & & \vdots \\ \varphi_{n1}(x) & \cdots & \varphi_{nn}(x) \end{pmatrix}.$$

Multiply the  $k$ th row ( $k \neq i$ ) of the left matrix by  $-a_{ik}(x)$  and then add it to the  $i$ th row. This process does not change the value of the determinant  $W_i(x)$ , but gives the relation

$$W_i(x) \equiv \det \begin{pmatrix} \varphi_{11}(x) & \cdots & \varphi_{1n}(x) \\ \vdots & & \vdots \\ a_{ii}\varphi_{i1}(x) & \cdots & a_{ii}\varphi_{in}(x) \\ \vdots & & \vdots \\ \varphi_{n1}(x) & \cdots & \varphi_{nn}(x) \end{pmatrix} = a_{ii}(x)W(x). \quad (16.43)$$

From (16.42) and (16.43), we arrive at the desired result. ♣

## 16.3 Autonomous Systems of ODEs

### 16.3.1 Autonomous System

We noted earlier that an  $n$ th-order ODE reduces to the first-order form:

$$\mathbf{y}(x)' \equiv \frac{d}{dx} \begin{bmatrix} y_1(x) \\ y_2(x) \\ \vdots \\ y_n(x) \end{bmatrix} = \begin{bmatrix} F_1(x; y_1, y_2, \dots, y_n) \\ F_2(x; y_1, y_2, \dots, y_n) \\ \vdots \\ F_n(x; y_1, y_2, \dots, y_n) \end{bmatrix} \equiv \mathbf{F}(x, \mathbf{y}), \quad (16.44)$$

where  $\mathbf{y}(x)$  and  $\mathbf{F}(x, \mathbf{y})$  are  $n$  column vectors. Particularly important in many applications is the case where  $\mathbf{F}(x, \mathbf{y})$  does not depend explicitly on  $x$ . Relevant terminology is given below.

#### ♠ Autonomous system of ODEs:

A system of a first-order ODE of the form

$$\mathbf{y}'(x) = \mathbf{F}(\mathbf{y})$$

is called an **autonomous system**, wherein  $\mathbf{F}$  does not depend explicitly on the independent variable  $x$ .

If  $\mathbf{F}$  does depend explicitly on  $x$ , the system is said to be **nonautonomous**.

*Example* Consider a second-order ODE of the form

$$u''(x) = f(u(x), u'(x)).$$

Setting  $y_1(x) = u(x)$  and  $y_2(x) = u'(x)$ , we have an autonomous system such as

$$\mathbf{y}'(x) = \frac{d}{dx} \begin{bmatrix} y_1(x) \\ y_2(x) \end{bmatrix} = \begin{bmatrix} y_2(x) \\ f(y_1(x), y_2(x)) \end{bmatrix} \equiv \mathbf{F}(\mathbf{y}).$$

### 16.3.2 Trajectory

As a prototype of autonomous systems of ODEs, we consider a two-dimensional system such that

$$\mathbf{y}'(t) = \frac{d}{dt} \begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix} = \begin{bmatrix} f_1(y_1, y_2) \\ f_2(y_1, y_2) \end{bmatrix} = \mathbf{f}(\mathbf{y}), \quad (16.45)$$

where  $y_1(t)$ ,  $y_2(t)$  are unknown functions on  $t$  in some interval  $I$ . We assume that  $f_1(y_1, y_2)$  and  $f_2(y_1, y_2)$  are defined in some domain  $D$  and satisfy the Lipschitz condition on both  $y_1(t)$  and  $y_2(t)$ . If  $t_0$  is any real number and  $(y_{10}, y_{20}) \in D$  for any  $y_{10} \equiv y_1(t_0)$  and  $y_{20} \equiv y_2(t_0)$ , the above hypotheses guarantee the existence and uniqueness of solutions for (16.45),

$$y_1(t) = \varphi_1(t), \quad y_2(t) = \varphi_2(t),$$

satisfying the initial conditions

$$\varphi_1(t_0) = y_{10}, \quad \varphi_2(t_0) = y_{20}.$$

We now consider a subdomain  $R$  of  $D$  in which  $f_1(y_1, y_2)$  does not vanish. Then, we have in  $R$  the relation

$$\frac{dy_2}{dy_1} = \frac{dy_2}{dt} \frac{dt}{dy_1} = \frac{dy_2/dt}{dy_1/dt} = \frac{f_2(y_1, y_2)}{f_1(y_1, y_2)}, \quad (16.46)$$

which represents a direction field in  $(y_1, y_2)$ -plane as noted in Sect. 15.1.6. From the uniqueness theorem, there exists a unique integral curve of (16.46) in  $R$  satisfying the initial conditions. Such an integral curve on  $(y_1, y_2)$ -plane is called a **trajectory** of (16.45).

♠ **Theorem:** At most one trajectory passes through any point.

*Proof* This is obvious from the uniqueness of solutions. If not, two or more trajectories emerge from the crossing point chosen as an initial value point. ♣

*Remark.* When the vector field

$$\mathbf{v}(x, y) = \begin{bmatrix} x'_1 \\ x'_2 \end{bmatrix} = \begin{bmatrix} f_1(x_1, x_2) \\ f_2(x_1, x_2) \end{bmatrix}$$

describes the motion of a point in  $R$ , the domain  $R$  is called a **phase space** of the system (16.45).

### 16.3.3 Critical Point

Suppose that the autonomous system (16.45) has a time-independent solution expressed by

$$\boldsymbol{\varphi}(t) = \mathbf{c} \in D,$$

where  $\mathbf{c} = (c_1, c_2)$  is a constant vector. Then, no trajectory can pass through the point  $\mathbf{c}$  (see the theorem in Sect. 16.3.3). In addition, we obviously have

$$\boldsymbol{\varphi}'(t) = \mathbf{0} = \mathbf{f}(\mathbf{c}).$$

Conversely, if there exists a point  $\mathbf{c}$  in  $R$  for which  $\mathbf{f}(\mathbf{c}) = \mathbf{0}$ , then the functions  $\boldsymbol{\varphi}(t) = \mathbf{c}$  are solutions of (16.45). The point  $\mathbf{c}$  is said to be a **critical point** (or **singular point** or **point of equilibrium**).

#### ♠ Critical point:

Assume an autonomous system

$$\mathbf{y}'(x) = \mathbf{F}(\mathbf{y}) \quad \text{for } \mathbf{y} \in D. \quad (16.47)$$

Then, any point  $\mathbf{c} \in D$  that gives

$$\mathbf{F}(\mathbf{c}) = \mathbf{0}$$

is called a **critical point** of (16.47). Any other point in  $D$  is called a **regular point**.

### 16.3.4 Stability of a Critical Point

Let us discuss the **stability of a critical point** of an autonomous system (16.45) by analyzing trajectories of its solutions around the critical point.

We assume throughout that the function  $\mathbf{F}(\mathbf{y})$  is differentiable of the first order on  $D$ , which guarantees the existence and uniqueness of solutions of the initial value problem (16.45). Then, the solutions of (16.45) can be conveniently pictured as curves in the phase space.

Now we consider a solution  $\psi(x)$  of (16.45) that passes through the point  $\eta$  for  $x_0$ , where the distance between  $\eta$  and  $\mathbf{c}$  is small. Let us now follow the trajectory that starts at a point  $\eta$  different from  $\mathbf{y}_0$ , but near  $\mathbf{c}$ . If the resulting motion  $\psi$  remains close to the critical point  $\mathbf{c}$  for  $x \geq x_0$ , then the critical point is said to be **stable**, but if the solution  $\psi$  tends to return to the critical point  $\mathbf{c}$  as  $x$  increases to infinity, then the critical point is said to be **asymptotically stable**. Finally, if the solution  $\psi$  leaves every small neighborhood of  $\mathbf{c}$ , the critical point is said to be **unstable**. More precisely, we have the following definitions:

♠ **Stability of a critical point:**

Let  $\mathbf{c}$  be a critical point of the autonomous system  $\mathbf{y}'(x) = \mathbf{F}(\mathbf{y})$ , so that  $\mathbf{F}(\mathbf{c}) = \mathbf{0}$ . The critical point  $\mathbf{c}$  is called:

(i) **stable** when given a positive  $\varepsilon$ , there exists a  $\delta$  so small that

$$|\mathbf{y}(0) - \mathbf{c}| < \delta \Rightarrow |\mathbf{y}(x) - \mathbf{c}| < \varepsilon \text{ for all } x > 0;$$

(ii) **asymptotically stable** when for some  $\delta$ ,

$$|\mathbf{y}(0) - \mathbf{c}| < \delta \Rightarrow \lim_{x \rightarrow \infty} |\mathbf{y}(x) - \mathbf{c}| = 0;$$

(iii) **strictly stable** when it is stable and asymptotically stable;

(iv) **neutrally stable** when it is stable but not asymptotically stable; and

(v) **unstable** when it is not stable.

### 16.3.5 Linear Autonomous System

An autonomous system  $\mathbf{y}' = \mathbf{F}(\mathbf{y})$  is called **linear** if and only if all the elements  $F_i$  of  $\mathbf{F}$  are linear homogeneous functions of the  $y_k$ , so that

$$\frac{dy_k}{dx} = a_{i1}y_1 + \cdots + a_{in}y_n \quad (i = 1, \dots, n).$$

Hence, a linear autonomous system is just a (homogeneous) linear system of ODEs with *constant* coefficients. The analyses for linear systems are generally useful since we can always replace  $F_i(\mathbf{y})$  by the linear terms of their Taylor expansions about a point  $\mathbf{y} = \mathbf{y}_0$  for analyzing their local behavior.

We now discuss in detail the case  $n = 2$  of linear plane autonomous systems of the form  $\mathbf{y}' = \mathbf{A}\mathbf{y}$ . Any such system is expressed by

$$\frac{dx}{dt} = ax + by, \quad \frac{dy}{dt} = cx + dy, \quad (16.48)$$

where  $x = y_1$ ,  $y = y_2$ , and

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

with  $a, b, c, d$  being constants. Observe that the simultaneous linear equations

$$ax + by = cx + dy = 0$$

have no solution except

$$x = y = 0$$

unless  $\det \mathbf{A} = 0$ . We thus see that the origin is the only critical point of the system (16.48) unless  $ad = bc$ .

Relevant terminology is given below.

♠ **Secular equation:**

If  $(x(t), y(t))$  is a solution of (16.48), then  $x(t)$  and  $y(t)$  satisfy the equation:

$$u'' - (a + d)u' + (ad - bc)u = 0. \quad (16.49)$$

This equation is called the **secular equation** of the autonomous system (16.48).

*Proof* The first equation of (16.48) says that

$$by = x' - ax,$$

which implies that

$$x'' - ax' = by'.$$

We thus have

$$x'' - ax' = b(cx + dy) = bcx + d(x' - ax),$$

or equivalently,

$$x'' - (a + d)x' + (ad - bc)x = 0.$$

The proof for  $y(t)$  is the same, replacing  $a$  with  $d$  and  $b$  with  $c$ . ♣

The secular equation (16.49) has an important property associated with the nature of the critical point. This is seen by introducing the concept of the **characteristic polynomial**  $P$  of (16.49) as

$$P \equiv \lambda^2 - (a + d)\lambda + (ad - bc) = \begin{vmatrix} a - \lambda & b \\ c & d - \lambda \end{vmatrix} = \det(\mathbf{A} - \lambda \mathbf{I}).$$

If  $\lambda_j$  ( $j = 1, 2$ ) are the roots of  $P = 0$ , then there exist nonzero **eigenvectors**  $(x_j, y_j)$  such that

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x_j \\ y_j \end{pmatrix} = \begin{pmatrix} ax_j + by_j \\ cx_j + dy_j \end{pmatrix} = \lambda_j \begin{pmatrix} x_j \\ y_j \end{pmatrix}.$$

From this, it follows that the functions

$$e^{\lambda_1 t} \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} \quad \text{and} \quad e^{\lambda_2 t} \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}$$

are a basis of vector-valued solutions of (16.48). We shall see later that the nature of a critical point of a system is completely determined by the values of the roots  $\lambda_1, \lambda_2$ .

## 16.4 Classification of Critical Points

The behavior of trajectories of a linear autonomous system

$$\frac{d}{dt} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \mathbf{A} \begin{pmatrix} u \\ v \end{pmatrix} \quad (16.50)$$

near its critical point depends on the eigenvalues of the matrix  $\mathbf{A}$ , denoted by  $\lambda_1$  and  $\lambda_2$ . There are five cases to consider and we discuss each in turn.

### 16.4.1 Improper Node

We first consider the case where  $\lambda_1$  and  $\lambda_2$  are *real, unequal, and of the same sign*. A critical point for this case is called an **improper node**. In this case, all the trajectories approach the critical point tangentially to the same straight line with increasing  $t$ .

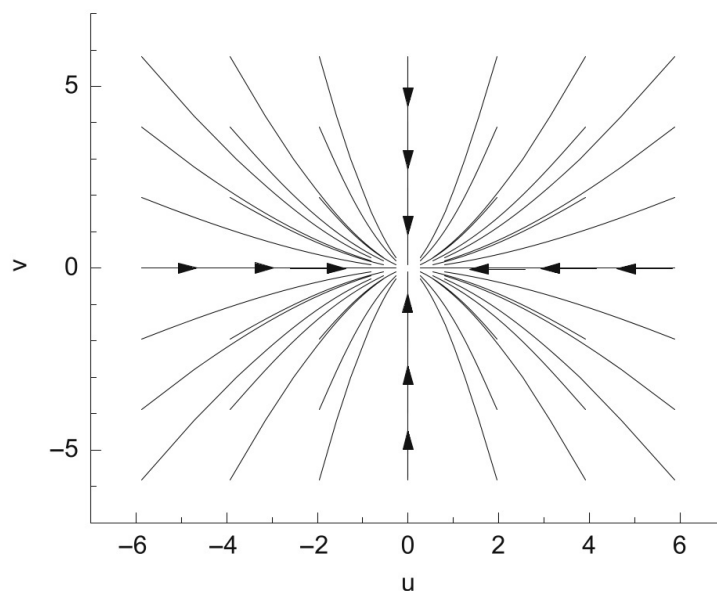
*Example* An example of improper nodes is given by

$$\frac{d}{dt} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} -2 & 0 \\ 0 & -3 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix}. \quad (16.51)$$

The eigenvalues are obviously  $\lambda = -3, -2$  and the corresponding eigenvectors are  $(1, 0)$  and  $(0, 1)$ . The general solution to (16.51) is

$$\begin{pmatrix} u \\ v \end{pmatrix} = c_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} e^{-2t} + c_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix} e^{-3t}. \quad (16.52)$$

The trajectories given by (16.52) for several values of  $c_1$  and  $c_2$  are shown in Fig. 16.1.



**Fig. 16.1.** Trajectories associated with the improper node of the system (16.51)

**Remark.** If the eigenvalues are real, unequal, and *positive* (contrary to the above example), then the trajectories are similar to those in Fig. 16.1 except that the directions of the arrows are reversed; in other words the trajectories recede from the critical point and go off toward infinity.

### 16.4.2 Saddle Point

We next consider the case where  $\lambda_1$  and  $\lambda_2$  are *real*, *unequal*, and of *the opposite sign*. In this case, the trajectories approach the critical point along one eigenvector direction and recede along the other eigenvector direction. The critical point in this case is called a **saddle point**.

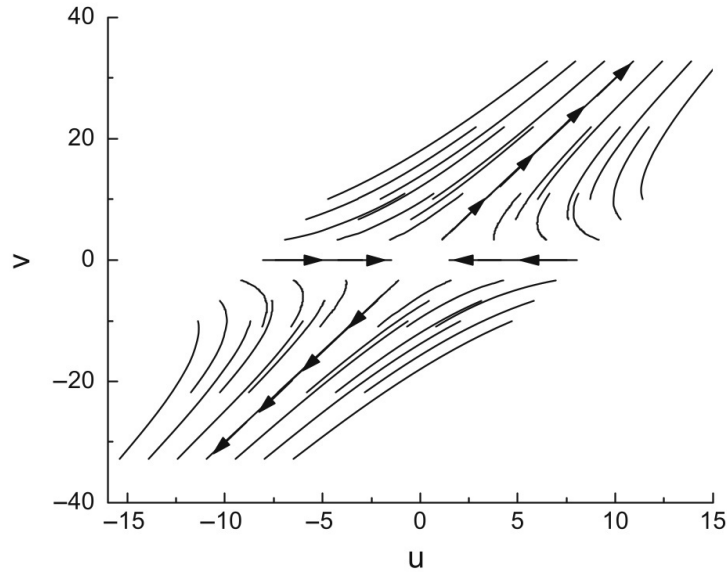
*Example* Assume the system

$$\frac{d}{dt} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} -1 & 1 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix}. \quad (16.53)$$

The eigenvalues are  $\lambda_{\pm} = -1, 2$ , and the corresponding eigenvectors are  $(1, 0)$  and  $(1, 3)$ , respectively. The general solution to (16.53) is

$$\begin{pmatrix} u \\ v \end{pmatrix} = c_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} e^{-t} + c_2 \begin{pmatrix} 1 \\ 3 \end{pmatrix} e^{2t}. \quad (16.54)$$

The trajectories given by (16.54) are shown in Fig. 16.2. As (16.54) consists of an  $e^{-t}$  term and an  $e^{2t}$  term, the trajectories approach the origin along the eigenvector direction  $(1, 0)$  and recede along the direction  $(1, 3)$  as  $t$  increases.



**Fig. 16.2.** Trajectories around the saddle point of the system (16.53)

### 16.4.3 Proper Node

We next consider the case of two roots of the characteristic equation being *real* and *equal*. This type of critical point is called a **proper node**.

*Example* We consider the critical point of the system

$$\frac{d}{dt} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \alpha & 0 \\ 0 & \alpha \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix}. \quad (16.55)$$

The critical point occurs at the origin, with the degenerate eigenvalue being  $\alpha$ . Generally when the eigenvalue  $\lambda$  of the characteristic equation is degenerate, the eigenvector is given by

$$u(t) = (c_1 + c_2 t)e^{\lambda t}, \quad v(t) = (c_3 + c_4 t)e^{\lambda t}. \quad (16.56)$$

Hence, we set  $\lambda = \alpha$  in (16.56) and substitute the results into (16.55) to obtain  $c_2 = c_4 = 0$ . The solution to (16.55) is thus

$$u(t) = c_1 e^{\alpha t}, \quad v(t) = c_3 e^{\alpha t}. \quad (16.57)$$

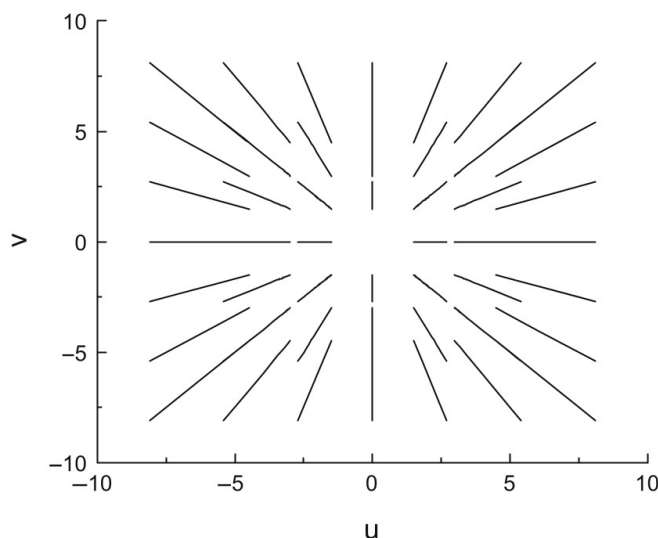
Eliminating  $t$  from (16.57) yields the expression of the trajectories:

$$v = \frac{c_3}{c_1} u \quad \text{if } c_1 \neq 0$$

and

$$u = 0 \quad \text{if } c_1 = 0,$$

both of which are depicted in Fig. 16.3. The trajectories approach or recede from the origin, depending on the sign of  $\alpha$ .



**Fig. 16.3.** Trajectories around the proper node of the system (16.55)

#### 16.4.4 Spiral Point

So far, we have restricted our attention to cases where the two eigenvalues are real. Now we consider the case in which the two eigenvalues are **complex conjugates** of each other. The corresponding critical point is called a **spiral point** or a **focus**.

*Example* An example for this case is

$$\frac{d}{dt} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 2 & -2 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix}. \quad (16.58)$$

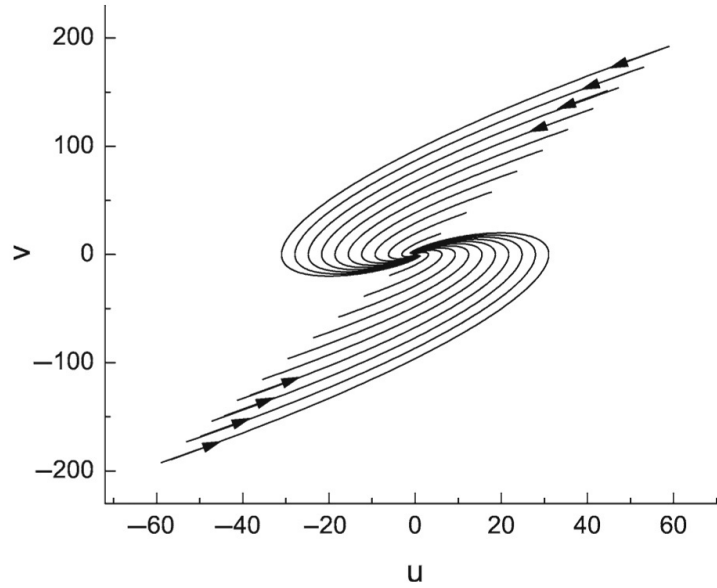
The critical point is at the origin, and the eigenvalues are  $\lambda_{\pm} = -1 \pm i$  with corresponding eigenvectors  $(1, 1 \mp i)$ . The general solution to this system is

$$\begin{pmatrix} u \\ v \end{pmatrix} = c_1 \begin{pmatrix} 1 \\ 1 - i \end{pmatrix} e^{(-1+i)t} + c_2 \begin{pmatrix} 1 \\ 1 + i \end{pmatrix} e^{(-1-i)t}. \quad (16.59)$$

The result represents a family of curves that spiral into the critical point as  $t$  increases. Real components of the solutions  $u(t)$  and  $v(t)$  given by (16.59) are plotted in Fig. 16.4.

#### 16.4.5 Center

The final class of critical points is called a **center**, for which the two eigenvalues are *pure imaginary*. In this case, trajectories consist of a family of closed loops centered about the critical point.



**Fig. 16.4.** Trajectories for the spiral point of the system (16.58)

*Example* Consider the system

$$\frac{d}{dt} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ -1 & -1 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix}. \quad (16.60)$$

The eigenvalues and corresponding eigenvectors are  $\lambda_{\pm} = \pm i$  and  $(1 \pm i, -1)$ , and the general solution reads

$$\begin{pmatrix} u \\ v \end{pmatrix} = c_1 \begin{pmatrix} 1 + i \\ -1 \end{pmatrix} e^{it} + c_2 \begin{pmatrix} 1 - i \\ -1 \end{pmatrix} e^{-it}. \quad (16.61)$$

Figure 16.5 shows several trajectories for different values of  $c_1$  and  $c_2$ . All the trajectories represent periodic motion about the critical point.

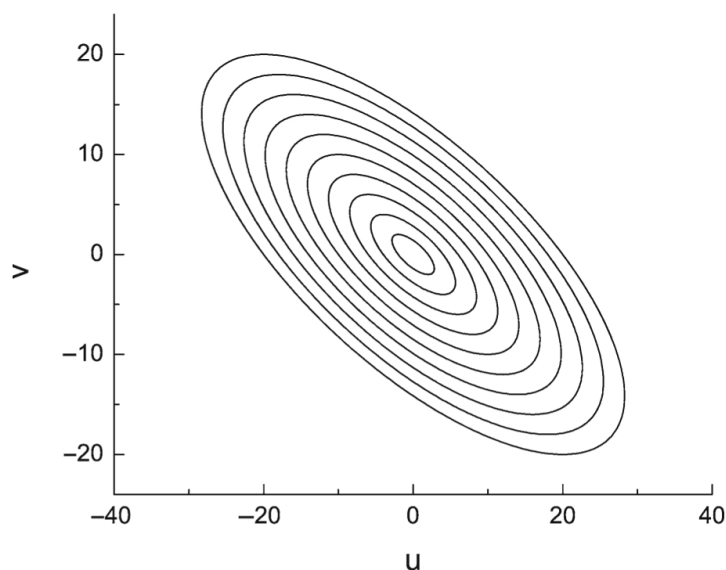
#### 16.4.6 Limit Cycle

Before closing this section, we have one more topic to discuss. Consider the system

$$\begin{aligned} x' &= x + y - x(x^2 + y^2), \\ y' &= -x + y - y(x^2 + y^2). \end{aligned} \quad (16.62)$$

The only critical point is at the origin. Letting  $x = r \cos \theta$  and  $y = r \sin \theta$ , the system (16.62) becomes

$$r' = r(1 - r^2) \quad \text{and} \quad \theta' = -1. \quad (16.63)$$

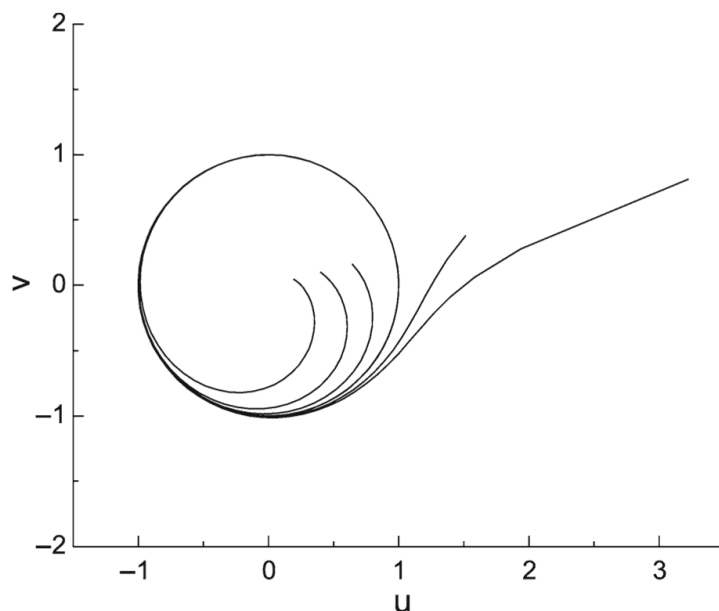


**Fig. 16.5.** Periodic trajectories for the **center** of the system (16.60)

The system (16.63) has a trivial solution of  $r = 1$ ,  $\theta = -t + \text{const}$ , which represents periodic clockwise motion around the unit circle. We can find other solutions by solving (16.63). The equation for  $r$  is easy to solve, yielding

$$r(t) = \frac{1}{\sqrt{1 + \left(\frac{1 - r_0^2}{r_0^2}\right) e^{-2t}}},$$

where  $r(0) = r_0$ . Figure 16.6 shows  $r(t)$  plotted against  $t$  for  $r_0 > 1$  and  $r_0 < 1$ . The trajectories spiral in toward the unit circle as  $t \rightarrow \infty$  if  $r_0 > 1$



**Fig. 16.6.** Converging behavior of solutions of the system (16.62) to a limit cycle

and they spiral out toward the unit circle as  $t \rightarrow \infty$  if  $r_0 < 1$ . Hence, all the trajectories spiral into the unit circle as  $t \rightarrow \infty$ .

The unit circle mentioned above is called a **limit cycle**. Limit cycles are important for determining the stability of the system, since the existence of a limit cycle ensures the existence of periodic solutions to a system.

## Exercises

1. Consider the system given by

$$\begin{aligned}x' &= e^x + \sin 5y - \cos 2y, \\y' &= x + 2 \sin y.\end{aligned}$$

Find the equilibrium point and describe the stability of the system around the point.

**Solution:** Expanding all functions on the right-hand side around  $x = 0$ ,  $y = 0$ , we have

$$\begin{aligned}x' &= x + 5y + g(x, y), \\y' &= x + 2y + h(x, y).\end{aligned}$$

The functions  $g, h$  converge to zero faster than  $\sqrt{x^2 + y^2}$ , and the characteristic equation becomes

$$\begin{vmatrix} -\lambda + 1 & 5 \\ 1 & -\lambda + 2 \end{vmatrix} = 0,$$

whose roots are  $(3 \pm \sqrt{21})/2$ . Both of these are positive and the system is unstable. ♣

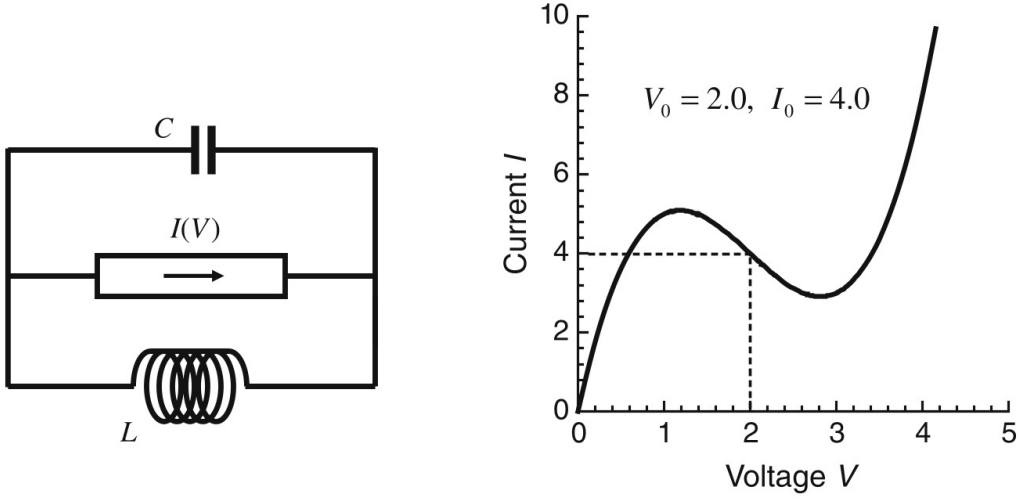
## 16.5 Applications in Physics and Engineering

### 16.5.1 Van der Pol Generator

As a physical example of a system in which a **limit cycle** may occur, we consider the following electric circuit consisting of a coil with inductance  $L$  and a condenser with capacitance  $C$  attached to a **tunnel diode**. A tunnel diode is a nonlinear element in the sense that it exhibits nonlinear current–voltage characteristics:

$$I(V) = I_0 - a(V - V_0) + b(V - V_0)^3. \quad (16.64)$$

It follows from Fig. 16.7 that a tunnel diode behaves like an ordinary resistor at low and high voltages, but like a *negative* resistor at intermediate voltages.



**Fig. 16.7.** *Left:* An electric circuit consisting of a coil with inductance  $L$ , a condenser with capacitance  $C$ , and a tunnel diode. *Right:* Plot of the nonlinear current–voltage characteristics  $I(V)$  of the tunnel diode

Thus, a tunnel diode is expected to amplify small oscillations in the system, provided we choose the parameters in an appropriate manner.

The equation of motion for the LC circuit attached to a tunnel diode is obtained as described below. The **law of the conservation current flow** ensures that

$$I_L + I(V) + I_C = 0, \quad (16.65)$$

where

$$I_L = \frac{1}{L} \int V dt \quad \text{and} \quad I_C = C \frac{dV}{dt}. \quad (16.66)$$

Substituting (16.64) and (16.66) into (16.65) and then differentiating with respect to time, we get

$$\frac{d^2 V}{dt^2} + \frac{1}{C} [-a + 3b(V - V_0)^2] \frac{dV}{dt} + \omega_0^2 V = 0,$$

where we introduce the resonant frequency  $\omega_0$  defined by the equation  $\omega_0^2 = 1/(LC)$ . For simplicity, we define a new variable

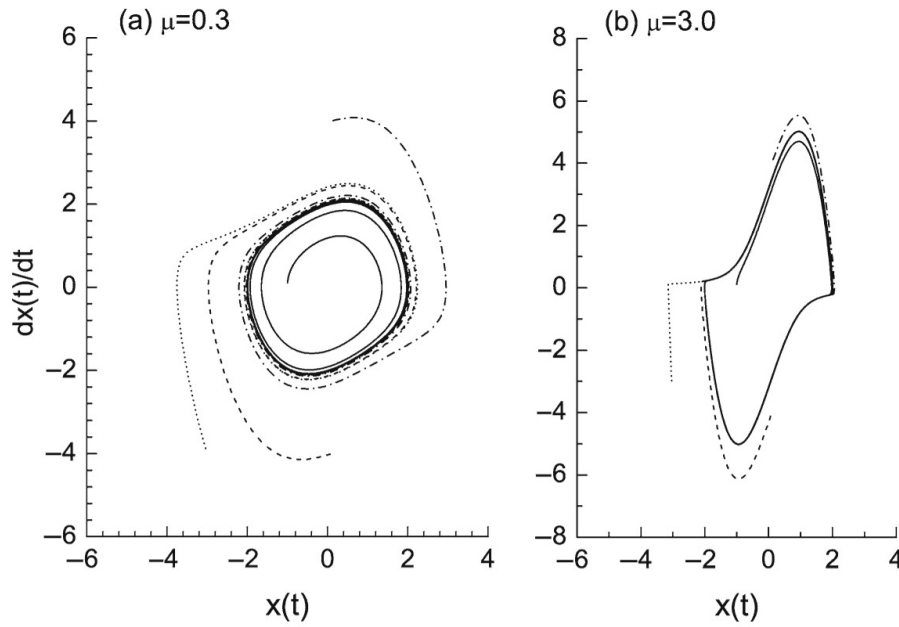
$$x = \frac{V - V_0}{V},$$

for which

$$\ddot{x} - \alpha(1 - \beta x^2)\dot{x} + \omega_0^2 x = 0 \quad \left( \dot{x} \equiv \frac{dx}{dt} \right) \quad (16.67)$$

with

$$\alpha = \frac{a}{C}, \quad \beta = \frac{3CV_0^2 b}{a}.$$



**Fig. 16.8.** Trajectories for the Van der Pol equation (16.68) with the parameter  $\mu = 0.3$  for (a) and  $\mu = 3.0$  for (b)

Equation (16.67) can be further simplified by replacing  $x$  by  $x/\sqrt{\beta}$  and introducing a new time variable  $\tilde{t} \equiv \omega_0 t$ . Hence, we finally obtain

$$\ddot{x} - \mu(1 - x^2)\dot{x} + x = 0 \quad (16.68)$$

with the following key parameter:

$$\mu = \frac{\alpha}{\omega_0} = a\sqrt{\frac{L}{C}}.$$

The nonlinear differential equation (16.68) is known as the **Van der Pol equation**. As shown below, it describes self-sustaining oscillations in which energy is supplied to small oscillations and removed from large oscillations, which gives rise to the limit cycle in the phase space.

We can observe the self-exciting behavior of the system governed by (16.68) in the phase space plot in Fig. 16.8, where we set  $\mu = 0.3$  and  $\mu = 3.0$  for various initial points  $\mathbf{x}_0 = (x(t=0), \dot{x}(t=0))$ . We see that all the trajectories starting at a point  $\mathbf{x}_0$  inside (or outside) a closed contour  $C$  move outward (or inward) as  $t$  increases and then converge to the contour  $C$ ; such a characteristic closed contour is known as the limit cycle of the system. The shape of the limit cycle depends on the value of  $\mu$ , as is evident from Fig. 16.8.

## Partial Differential Equations

---

**Abstract** Broadly speaking, there are three classes of partial differential equations that are relevant to mathematical physics, as reflected in the section titles of this chapter. After examining the basic properties common to all the abovementioned classes of equations, we devote the balance of this chapter to a discussion of the mathematical essence of each class.

### 17.1 Basic Properties

#### 17.1.1 Definitions

In this section we present the basic theory of **partial differential equations** (PDEs), an understanding of which is crucial is for describing or predicting the realm of nature. The formal definition is given below.

♠ **Partial differential equations:**

A partial differential equation of order  $r$  is a functional equation of the form

$$F\left(x_1, x_2, \dots, x_n; \frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}, \dots, \frac{\partial u}{\partial x_n}; \frac{\partial^2 u}{\partial x_1^2}, \dots\right) = 0, \quad (17.1)$$

which involves at least one  $r$ th-order partial derivative of the unknown function  $u = u(x_1, x_2, \dots, x_n)$  of independent variables  $x_1, x_2, \dots, x_n$ .

In this chapter we often denote partial derivatives with subscripts such as

$$\frac{\partial u}{\partial x} \equiv u_x \equiv \partial_x u, \quad \frac{\partial^2 u}{\partial x \partial y} \equiv u_{xy} \equiv \partial_x \partial_y u,$$

We also use the shorthand

$$\partial_j \equiv \frac{\partial}{\partial x_j}, \quad \partial_i \partial_j = \frac{\partial^2}{\partial x_i \partial x_j},$$

Then, the general form (17.1) of a PDE is expressed as

$$F(x, y, \dots, u, u_x, \dots, u_{xx}, u_{xy}, \dots) = 0, \quad (17.2)$$

where  $u = u(x, y, \dots)$  is the unknown function of independent variables  $x, y, \dots$ . A **solution** (or **integral**) of a PDE is a function  $\varphi(x, y, \dots)$  satisfying equation (17.2) identically, at least in some region of the independent variables  $x, y, \dots$ .

### 17.1.2 Subsidiary Conditions

The general solution of (17.1) depends on an arbitrary function. This statement is valid even for higher-order PDEs, indicating that a PDE has in general many solutions. Hence, in order to determine a unique solution, auxiliary conditions must be imposed. Such conditions are usually called **initial conditions** on time or **boundary conditions** for positions.

#### Initial condition:

In physics, an unknown function in a PDE usually involves independent variables of time  $t$  and position  $x, y, \dots$ . Initial conditions for an unknown function are imposed on a particular (initial) time  $t = t_0$  for an unknown function and/or its time derivatives.

#### Boundary condition:

Boundary conditions are imposed for an unknown function at the boundary or the infinity of a domain  $D$  in which the PDE is valid and are classified into two cases:

1. **Dirichlet condition** is the case in which an unknown function  $u$  is specified on the boundary of the domain  $D$  (often denoted by  $\partial D$ ), where  $u$  is a function of time  $t$  and **position**  $x, y, \dots$ .
2. **Neumann condition** is the case in which the normal derivative of an unknown function  $\partial u / \partial n$  is specified.

### 17.1.3 Linear and Homogeneous PDEs

A PDE is called **linear** if and only if the  $F$  of (17.1) is a linear function of  $u$  and its derivatives. First we assume a first-order PDE with two independent variables  $x$  and  $y$ , whose general form reads

$$F(x, y; u, u_x, u_y) = 0. \quad (17.3)$$

Then, if it is linear, (17.3) can be expressed by

$$a(x, y)u_x + b(x, y)u_y + c(x, y)u = g(x, y), \quad (17.4)$$

where  $a$ ,  $b$ ,  $c$ , and  $g$  are given functions of  $x$ ,  $y$ . Using the operator  $L$ , we express (17.4) by a simple form such that

$$Lu(x, y) = g(x, y), \quad (17.5)$$

where the operator  $L$  is defined by

$$L = a(x, y)\partial_x + b(x, y)\partial_y + c(x, y).$$

The linearity of PDEs guarantees that for any function  $u$ ,  $v$  and any constant  $c$  the relations hold for

$$L(u + v) = Lu + Lv, \quad L(cu) = cL(u).$$

### Examples

$$u_{xx} - e^{-x}u_{yy} = 0 \quad (\text{linear})$$

$$u_{xx} - e^{-x}u_{yy} = \sin x \quad (\text{linear})$$

$$uu_x + u_y = 0 \quad (\text{nonlinear})$$

$$xu_x + yu_y + u^2 = 0 \quad (\text{nonlinear})$$

A linear equation is said to be **homogeneous** if the equation contains either the dependent variable  $u$  or its derivatives  $u_x, u_y, \dots$ , not an independent variable such as  $x, y, \dots$ . For instance, the PDE (17.5) is homogeneous if

$$g(x, y) \equiv 0,$$

since the equation

$$Lu(x, y) = 0 \quad (17.6)$$

involves only  $u, u_x, u_y$  and not  $x$  or  $y$ . On the other hand, if  $g \neq 0$  in (17.5), it is called an **inhomogeneous** (or **nonhomogeneous**) linear equation. These statements are generally valid even for higher-order PDEs.

#### 17.1.4 Characteristic Equation

We consider a **first-order homogeneous linear** PDE of the form

$$a(x, y)\partial_x u(x, y) + b(x, y)\partial_y u(x, y) = 0, \quad (17.7)$$

which is the most simple (and thus pedagogical) class of PDEs. In general, solutions of PDEs are described by *arbitrary* functions  $f(p)$  of a particular independent variable  $p$ , wherein  $p = p(x, y)$  is some combination of independent variables  $x$  and  $y$ . We verify this statement for the case of (17.7).

By the chain rule on the derivative, we have

$$\frac{\partial u}{\partial x} = \frac{\partial p}{\partial x} \frac{df}{dp}, \quad \frac{\partial u}{\partial y} = \frac{\partial p}{\partial y} \frac{df}{dp}. \quad (17.8)$$

Hence, the PDE (17.7) can be rewritten in the form

$$\left[ a(x, y) \frac{\partial p}{\partial x} + b(x, y) \frac{\partial p}{\partial y} \right] \frac{df}{dp} = 0. \quad (17.9)$$

This implies that the function form of  $f(p)$  may be arbitrary if  $p = p(x, y)$  satisfies the equation

$$a(x, y) \frac{\partial p}{\partial x} + b(x, y) \frac{\partial p}{\partial y} = 0. \quad (17.10)$$

Therefore, an arbitrary function  $f(p)$  such that the  $p$  satisfies (17.10) serves as the solution of the original PDE of (17.7). [The case of  $\partial f / \partial p = 0$  gives a trivial solution of  $f(p) = u(x, y) = \text{const}$ , which we omit below.]

To obtain the solution  $p = p(x, y)$  of the equation (17.10), we tentatively suppose that the function  $p = p(x, y)$  takes a constant value along a curve  $C : y = y(x)$  on the  $(x, y)$ -plane. Then, the total derivative of  $p$  on the curve  $C$  should vanish, so that

$$dp = \frac{\partial p}{\partial x} dx + \frac{\partial p}{\partial y} dy = 0. \quad (17.11)$$

From the correspondence between (17.10) and (17.11), we see that these are equivalent provided that

$$\frac{dy}{dx} = \frac{b(x, y)}{a(x, y)}, \quad a(x, y) \neq 0. \quad (17.12)$$

This is called the **characteristic equation** of the PDE (17.7) and its solution  $y = y(x)$  is the **characteristic curve** of (17.7). From (17.11) and (17.12), therefore, we obtain the desired function form of  $p = p(x, y)$  that makes an arbitrary function  $f(p)$  the solution of the original PDE.

*Examples* We evaluate a general solution for (17.7) in the case that  $a, b$  are constant and nonzero coefficients. From (17.11) and (17.12), we obtain the characteristic curve (line)  $p = bx - ay$ . Then a general solution takes the form

$$u(x, y) = f(p) = f(bx - ay), \quad (17.13)$$

where  $f$  is an arbitrary function. The solution can be easily checked by taking derivatives using (17.8) and substituting those into (17.7). A less trivial example is given in Exercise 1.

### 17.1.5 Second-Order PDEs

The general form of second-order linear PDEs is

$$\sum_{i,j=1}^n a_{ij}(x_1, x_2, \dots, x_n) \partial_i \partial_j u + \sum_{j=1}^n a_i(x_1, x_2, \dots, x_n) \partial_i u + a_0(x_1, x_2, \dots, x_n) u = g(x_1, x_2, \dots, x_n), \quad (17.14)$$

where the unknown function  $u$  depends on  $n$ -independent variables denoted by  $x_1, x_2, \dots, x_n$ . Note that  $a_{ij} = a_{ji}$  since the mixed derivatives are equal. The form of (17.14) represents a very large class of PDEs. Among them, we restrict our attention to the case  $g = 0$  with real constant coefficients, namely, second-order linear homogeneous PDEs. The general form of linear PDEs of second-order involving  $n$  independent variables with real constant coefficients is written as

$$\sum_{i,j=1}^n a_{ij} \partial_i \partial_j u + \sum_{j=1}^n a_i \partial_i u + a_0 u = 0. \quad (17.15)$$

The linear transformation of independent variables  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  to  $\mathbf{y} = (y_1, y_2, \dots, y_n)$  is given by

$$\mathbf{y} = B\mathbf{x}, \quad (17.16)$$

or equivalently,

$$y_k = \sum_{m=1}^n b_{km} x_m,$$

where the  $b_{km}$  are elements of the  $n \times n$  matrix  $B$ . Using the chain rule on the derivative, we have

$$\frac{\partial}{\partial x_j} = \sum_{k=1}^n \frac{\partial y_k}{\partial x_j} \frac{\partial}{\partial y_k}$$

and

$$\frac{\partial^2 u}{\partial x_i \partial x_j} = \left( \sum_{k=1}^n b_{ki} \frac{\partial}{\partial y_k} \right) \left( \sum_{m=1}^n b_{mj} \frac{\partial}{\partial y_m} \right) u. \quad (17.17)$$

Hence, the first term of (17.15) is converted to

$$\sum_{i,j=1}^n a_{ij} \partial_i \partial_j u = \sum_{k,m=1}^n (b_{ki} a_{ij} b_{mj}) \partial_k \partial_m u,$$

which leads the relation

$$a_{ij} \rightarrow \sum_{k,m=1}^n b_{ki} a_{ij} b_{mj}. \quad (17.18)$$

Thus we obtain the PDE with new variables  $y_1, y_2, \dots, y_n$  by the transformation  $A \rightarrow B^t AB$ , where  $B^t$  is the transpose of  $B$ .

The appropriate choice of the matrix  $B$  makes it possible to diagonalize  $A$  such that

$$B^t AB = \begin{pmatrix} c_1 & & & \\ & c_2 & 0 & \\ & 0 & \ddots & \\ & & & c_n \end{pmatrix},$$

where  $c_1, c_2, \dots, c_n$  are the real eigenvalues of the matrix  $A$ . Thus, any PDE of the form (17.15) can be converted into a PDE with diagonal coefficients in terms of a linear transformation of a set of independent variables such as

$$\sum_{i=1}^n c_i \frac{\partial^2 u}{\partial y_i^2} + \sum_{i=1}^n d_i \frac{\partial u}{\partial y_i} = 0. \quad (17.19)$$

♠ **Theorem:**

By linear transformation of independent variables, the equation (17.15) can be reduced to the canonical form (17.19).

### 17.1.6 Classification of Second-Order PDEs

We can classify the types of PDEs depending on the positive or negative values of the coefficients  $c_1, c_2, \dots, c_n$  in (17.19) for the case  $d_i \equiv 0$ .

1. **Elliptic case:**

If all the eigenvalues  $c_1, c_2, \dots, c_n$  are positive or negative, the PDE is called elliptic. A simple example is given by

$$\frac{\partial^2 u}{\partial y_1^2} + \frac{\partial^2 u}{\partial y_2^2} + \dots = 0.$$

2. **Hyperbolic case:**

In this case none of the  $\{c_i\} : i = 1, 2, \dots, n$  vanish and one of them has the opposite sign from  $n - 1$  than the others. For example,

$$\frac{\partial^2 u}{\partial y_1^2} - \frac{\partial^2 u}{\partial y_2^2} + \dots = 0.$$

3. **Parabolic case:**

If one of the  $\{c_i\}$ ,  $i = 1, 2, \dots, n$  is zero and all the others have the same sign, the PDE is parabolic.

Below are basic PDEs in physics classified by the definition given above:

$$\textbf{Laplace equation: } \Delta_n u = 0, \quad (17.20)$$

$$\textbf{Wave function: } u_{tt} = \Delta_n u, \quad (17.21)$$

Here  $\Delta_n$  means the **Laplacian** defined by  $\Delta_n = \partial_1^2 + \partial_2^2 + \cdots + \partial_n^2$ . The other important equation takes the form

$$u_t = \Delta_n u, \quad (17.22)$$

which we call the **diffusion equation**. The diffusion equation is different from the wave equation, where the time reversal symmetry  $t \rightarrow -t$  holds. All of these equations (17.20)–(17.22) are **linear** since they are first degree in the dependent variable  $u$ .

## Exercises

1. Find a general solution of the PDE of  $u = u(x, y)$  given by

$$u_x + 2xy^2 u_y = 0. \quad (17.23)$$

**Solution:** The characteristic equation of (17.23) reads  $dy/dx = 2xy^2$ , which has the solution  $y = 1/(p - x^2)$ . Hence, we have  $p = x^2 + (1/y)$ , i.e., the general solution is given by

$$u(x, y) = f\left(x^2 + \frac{1}{y}\right).$$

In fact,  $u(x, y)$  is a constant on the characteristic curve  $y = 1/(p - x^2)$  whatever value  $p$  takes, as proved by

$$\frac{d}{dx} u\left(x, \frac{1}{p - x^2}\right) = \frac{\partial u}{\partial x} + \frac{2x}{(p - x^2)^2} \frac{\partial u}{\partial y} = \frac{\partial u}{\partial x} + 2xy^2 \frac{\partial u}{\partial y} = 0,$$

and similarly we have  $du/dy = 0$ . ♣

2. Classify second-order PDEs in two independent variables whose general form is given by

$$\partial_x^2 u + 2a_{12} \partial_x \partial_y u + a_{22} \partial_y^2 u = 0, \quad (17.24)$$

where  $a_{12}$ ,  $a_{22}$  are real constants.

**Solution:** By completing the square, we can write (17.24) as

$$(\partial_x + a_{12} \partial_y)^2 u + (a_{22} - a_{12}^2) \partial_y^2 u = 0. \quad (17.25)$$

Here, let us introduce the new variables  $z$  and  $w$  by the linear transformation of the form  $x = z$ ,  $y = a_{12}z + (a_{22} - a_{12}^2)^{1/2}w$ . We then have

$$\frac{\partial}{\partial z} = \frac{\partial}{\partial x} + a_{12} \frac{\partial}{\partial y}, \quad \frac{\partial}{\partial w} = (a_{22} - a_{12}^2)^{1/2} \frac{\partial}{\partial y},$$

so that for the case  $a_{22} > a_{12}^2$  (17.25) gives

$$\frac{\partial^2 u}{\partial z^2} + \frac{\partial^2 u}{\partial w^2} = 0.$$

This is the elliptic case and is called the Laplace equation in the  $zw$ -plane. We easily see that for (17.25) the hyperbolic case is obtained for  $a_{22} < a_{12}^2$ . Thus, the second term of (17.25) determines the types of PDEs. ♣

## 17.2 The Laplacian Operator

### 17.2.1 Maximum and Minimum Theorem

We describe the fundamental properties of three operators, the Laplace, diffusion, and wave operators. These three operators are of great importance in the theory of PDEs. We begin with a description of the **Laplace operator** (or simply **Laplacian**)  $\Delta_n$  on  $\mathbf{R}^n$  defined by

$$\Delta_n = \sum_{i=1}^n \partial_i^2,$$

where  $n$  is a positive integer. The Laplacian is not only important in its own right, but also forms the spatial component of the **diffusion operator**  $L_D = \partial_t - \Delta_n$  and the **wave operator**  $L_W = \partial_t^2 - \Delta_n$ , whose properties are discussed in Sect. 17.3 and 17.4.

First, we explain the **maximum principle** for the **Laplace equation** given by

$$\Delta_n u(\mathbf{x}) = 0,$$

whose solutions are called **harmonic functions**. Obviously, the one-dimensional case ( $n = 1$ ) is trivial, so we consider the case where  $n > 1$ . Let  $D$  be a connected open set and  $u$  be an harmonic function in  $D$  with  $\sup u(\mathbf{x}) = A < \infty$  for  $\mathbf{x} \in D$ .

#### ♠ Maximum and minimum theorem:

The maximum and minimum values of  $u$  are achieved on  $\partial D$ , say the boundary of  $D$ , and nowhere inside.

Before going to the proof, we examine certain properties of the solutions of **Poisson's equation** expressed by

$$\Delta_n u(\mathbf{x}) = -4\pi\rho(\mathbf{x}). \quad (17.26)$$

♠ **Lemma:**

If the function  $\rho(\mathbf{x})$  in Poisson's equation (17.26) is positive (or negative) at a point  $\mathbf{x}_0$ , then the solution of (17.26) cannot attain its maximum (or minimum) value at the point  $\mathbf{x}_0$ .

**Proof (of the lemma):** If the function  $u(\mathbf{x})$  satisfying (17.26) attains a minimum at a point  $\mathbf{x}_0$ , then it should attain a minimum with respect to each component  $x_1, x_2, \dots, x_n$  separately at that point. Then all the second-order derivatives of  $u$  would have to be non-negative, which means that the left-hand side of (17.26), i.e., the sum of the second-order derivatives would have to be nonnegative. This result contradicts our hypothesis that  $\rho(\mathbf{x})$  in (17.26) is positive. Hence, the first part of the lemma has been proved. The second part of the lemma is proved in a similar manner by assuming that  $\rho(\mathbf{x})$  is negative. ♣

We are now ready to verify the maximum and minimum theorem.

**Proof (of the maximum and minimum theorem):** The proof is by contradiction. We first assume that

$$u(\mathbf{x}_0) > u_b + \varepsilon,$$

where  $u_b$  is the value of the function  $u(\mathbf{x})$  at an arbitrary point on the boundary of the defining domain  $D$ . We further assume the function

$$v(\mathbf{x}) = u(\mathbf{x}) + \eta r(\mathbf{x})^2,$$

where

$$r(\mathbf{x})^2 = |\mathbf{x} - \mathbf{x}_0|^2$$

and  $\eta$  is some positive constant. It then follows that

$$\Delta_n v = \Delta_n u + 2n\eta = 2n\eta,$$

which says that the  $v(\mathbf{x})$  is a solution of Poisson's equation (17.26) with negative  $\rho(\mathbf{x})$ . Note that  $v(\mathbf{x}_0) = u(\mathbf{x}_0)$  and by hypothesis

$$u(\mathbf{x}_0) > u_b + \varepsilon = v_b + \varepsilon - \eta r^2.$$

Choosing  $\eta$  to be so small that throughout  $D$

$$\varepsilon - \eta r^2 > \frac{\varepsilon}{2},$$

we obtain

$$v(\mathbf{x}_0) > v_b + \frac{\varepsilon}{2},$$

which implies that  $v$  attains its maximum somewhere within the domain  $D$ . This clearly contradicts the lemma above, so our assumption at the beginning of the proof was false. ♣

### 17.2.2 Uniqueness Theorem

The following theorem establishes the uniqueness of the solution of the Dirichlet problems for the Laplace equations.

♠ **Uniqueness theorem:**

If it exists, the solution of the Dirichlet problem for a Laplace equation is unique.

*Proof* Suppose that  $u_1$  and  $u_2$  are solutions on  $D$  for the Dirichlet problem such that

$$\Delta_n u = f(\mathbf{x}) \text{ in } D,$$

$$u = g(\mathbf{x}) \text{ on } \partial D.$$

Let  $w = u_1 - u_2$ , then  $\Delta_n w = 0$  in  $D$  and  $w = 0$  on  $\partial D$ . By the maximum (or minimum) principle, the point  $\mathbf{x}_m$  (or  $\mathbf{x}_M$ ) that minimizes (or maximizes)  $w(\mathbf{x})$  should be located on the boundary of  $D$ . Hence, we have

$$0 = w(\mathbf{x}_m) \leq w(\mathbf{x}) \leq w(\mathbf{x}_M) = 0$$

for all  $\mathbf{x} \in D$ , which means that  $w = 0$  and  $u_1 = u_2$ . ♣

### 17.2.3 Symmetric Properties of the Laplacian

The Laplacian is invariant under all rigid transformations such as translations and rotations. A translation from  $\mathbf{x}$  to a new variable  $\mathbf{x}'$  is given by

$$\mathbf{x}' = \mathbf{x} + \mathbf{a},$$

where  $\mathbf{a}$  is a constant vector in  $n$ -dimensional space. The rotation is expressed by

$$\mathbf{x}' = B\mathbf{x}, \tag{17.27}$$

where  $B$  is an orthogonal matrix with the property  $BB^t = B^tB = I$ . Invariance under translations or rotations means simply that

$$\sum_{i=1}^n \frac{\partial^2}{\partial x_i^2} = \sum_{j=1}^n \frac{\partial^2}{\partial x_j'^2}.$$

The proof for translational invariance is simple, so we leave it to the reader. In physical systems, translational invariance is apparent because the physical laws are independent of the choice of coordinates.

A rotational invariance under (17.27) is proved by using the chain rule on the derivative such that

$$\sum_{k=1}^n \frac{\partial^2}{\partial x_k^2} = \sum_{i,j} \sum_k b_{ik} b_{jk} \frac{\partial^2}{\partial x_i' \partial x_j'} = \sum_{i,j} \delta_{ij} \frac{\partial^2}{\partial x_i' \partial x_j'} = \sum_{i=1}^n \frac{\partial^2}{\partial x_i'^2},$$

where we have used the relation

$$\sum_k b_{ik} b_{jk} = (BB^t)_{ij} = \delta_{ij}.$$

Thus the proof has been completed.

Rotational invariance suggests that a two- or three-dimensional Laplacian should take a particularly simple form in polar or spherical coordinates.

## Exercises

1. Find the harmonic function for a two-dimensional Laplace equation that is invariant under rotations.

**Solution:** The two-dimensional Laplacian in polar coordinates is given by

$$\Delta_2 = \frac{\partial^2}{\partial r^2} + \frac{1}{r} \frac{\partial}{\partial r} + \frac{1}{r^2} \frac{\partial^2}{\partial \vartheta^2} \quad (r > 0),$$

where we seek for solutions  $u(r)$  depending only on  $r$ . Then we take the **radial part** of the Laplace equation, which gives  $u_{rr} + \frac{1}{r}u_r = 0$  ( $r > 0$ ). This is the ODE whose solution is given by  $u(r) = a \log r + b$  ( $r > 0$ ), where  $a, b$  are constants. Note that the form of the function  $\log r$  is scale invariant under the dilatation transformation  $r \rightarrow cr$  for a positive constant  $c$ . ♣

2. Find the harmonic function in three dimensions that is invariant under rotations.

**Solution:** The Laplacian in spherical coordinates takes the form

$$\Delta_3 = \frac{\partial^2}{\partial r^2} + \frac{2}{r} \frac{\partial}{\partial r} + \frac{1}{r^2 \sin \vartheta} \frac{\partial}{\partial \vartheta} \left( \sin \vartheta \frac{\partial}{\partial \vartheta} \right) + \frac{1}{r^2 \sin^2 \vartheta} \frac{\partial^2}{\partial \phi^2} \quad (r > 0).$$

Since the solution depends only on  $r$ , we have the Laplace equation given by  $u_{rr} + \frac{2}{r}u_r = 0$  ( $r > 0$ ). So we have  $(r^2 u_r)_r = 0$  and the solution becomes  $u = \frac{a}{r} + b$  ( $r > 0$ ), where  $a, b$  are constants. This is an important harmonic function that is not finite at the origin. ♣

3. Show that, for an arbitrary integer  $n > 2$ , the general form of solutions with rotational symmetry is given by

$$u(r) = ar^{2-n} + b \quad (n > 2, r > 0), \quad (17.28)$$

where  $a, b$  are constants.

**Solution:** This is shown by applying the chain rule to the derivative such that

$$\begin{aligned} \Delta_n u(r) &= \sum_{i=1}^n \partial_i \left[ \frac{x_i}{r} u'(r) \right] = \sum_{i=1}^n \left[ \frac{x_i^2}{r^2} u''(r) + \frac{1}{r} u'(r) - \frac{x_i^2}{r^3} u'(r) \right] \\ &= u''(r) + \frac{n-1}{r} u'(r), \end{aligned} \quad (17.29)$$

where the relation  $\partial r / \partial x_i = x_i / r$  is used. If  $\Delta_n u = 0$ , (17.29) yields

$$\frac{u''(r)}{u'(r)} = \frac{1-n}{r}.$$

Integrating twice, we have (17.28). ♣

## 17.3 The Diffusion Operator

### 17.3.1 The Diffusion Equations in Bounded Domains

The diffusion equation describes physical phenomena such as **Brownian motion** of a particle or **heat flow**, whose general form is written as

$$L_D u(x_1, x_2, \dots, t) = 0, \quad (17.30)$$

where  $L_D$  is the diffusion operator defined by

$$L_D = \partial_t - \sum_{i=1}^n \partial_i^2. \quad (17.31)$$

If the scale transformation  $t \rightarrow Dt$  is used, we have the **diffusion equation**

$$\partial_t u - D\Delta u = 0,$$

where  $D$  is the **diffusion constant**. For heat flow  $u$  represents the temperature at position  $\mathbf{x} = (x_1, x_2, \dots)$  and time  $t$ , and for Brownian motion  $u$  is the probability of finding a particle at  $\mathbf{x}$  and  $t$ . Hereafter, we treat the system of the unit diffusion constant  $D = 1$ . If we have to go back to the actual diffusion equation, we do the transformation  $t \rightarrow Dt$  in the final solution.

### 17.3.2 Maximum and Minimum Theorem

We begin by describing the maximum principle for the diffusion equation defined in a bounded domain, from which we deduce the uniqueness of initial and boundary value problems.

♠ **Maximum and minimum theorem:**

Let  $D$  be a bounded domain in  $\mathbf{R}^n$  and  $0 < t < T < \infty$ . If  $u$  is a real-valued continuous function, it takes its maximum either at the initial value ( $t = 0$ ) or on the boundary  $\partial D$ .

*Proof* For any  $\varepsilon > 0$ , we set

$$v(\mathbf{x}, t) = u(\mathbf{x}, t) + \varepsilon |\mathbf{x}|^2,$$

for which we have

$$v_t - \Delta v = -2n\varepsilon < 0. \quad (17.32)$$

If the maximum for  $u$  occurs at an interior point  $(\mathbf{x}_0, t_0)$  in the domain  $D \times [0, T]$ , we know that the first derivatives  $v_t, v_{x_1}, v_{x_2}, \dots$  of  $v$  vanish there and that the second derivative  $\Delta v \leq 0$ . This contradicts (17.32), so there is no interior maximum. Suppose now that the maximum occurs at  $t = T$  on  $D$ ; the time derivative  $v_t$  must be nonnegative there because

$$v(\mathbf{x}_0, T) \geq v(\mathbf{x}_0, T - \delta)$$

and

$$\Delta v \leq 0,$$

which again contradicts (17.32). Therefore, the maximum must be at the initial time  $t = 0$ , namely,  $D \times \{0\}$  or the boundary  $\partial D$ . Replacing  $u$  by  $-u$ , we see that the minimum is also achieved on either  $D \times \{0\}$  or  $\partial D$ . ♣

### 17.3.3 Uniqueness Theorem

The maximum principle can be used to prove **uniqueness** for the **Dirichlet problem for the diffusion equation**. The conditions are given by

$$\begin{aligned} L_D u &= f(\mathbf{x}, t) \quad \text{on } D \times (0, \infty) \\ u(\mathbf{x}, 0) &= g(\mathbf{x}), \quad u(\mathbf{x}, t) = h(t) \quad \text{on } \partial D \end{aligned}$$

for given functions  $f, g$ , and  $h$ .

The following is an immediate corollary of the maximum and minimum theorem.

♠ **Uniqueness theorem:**

There is at most one solution of the Dirichlet problem for the diffusion equation.

*Proof* Let  $u(\mathbf{x}, t)$  and  $v(\mathbf{x}, t)$  be two solutions of (17.33) and  $w = u - v$  be their difference. Hence, we have  $L_D w = 0$ ,  $w(\mathbf{x}, 0) = 0$ ,  $w(0, t) = 0$ ,  $w(\mathbf{x}, t) = 0$  on  $\partial D$ . By the maximum principle,  $w(\mathbf{x}, t)$  has its maximum at the initial time or the boundary, exactly where  $w$  vanishes. Thus  $w(\mathbf{x}, t) \leq 0$ . The same reasoning for the minimum shows that  $w(\mathbf{x}, t) \geq 0$ . Therefore  $w(\mathbf{x}, t) = 0$ , so that  $u = v$  for all  $t \geq 0$ . ♣

## 17.4 The Wave Operator

### 17.4.1 The Cauchy Problem

The **wave operator** (or **d'Alembertian**) on  $\mathbf{R}^n \times \mathbf{R}$  is expressed by

$$L = \partial_t^2 - \Delta_n = \partial_t^2 - \sum_{i=1}^n \partial_i^2, \quad (17.33)$$

from which we have the wave equation in the general form

$$Lu = \partial_t^2 u - \Delta_n u = 0. \quad (17.34)$$

The wave equation is the prototype of the hyperbolic PDEs and describes waves with unit velocity of propagation in homogeneous isotropic media. By making the transformation  $t \rightarrow ct$ , we have the standard form of the wave equation

$$\partial_t^2 u - c^2 \Delta_n u = 0, \quad (17.35)$$

where  $c$  is the wave velocity. The solution for (17.35) is obtained by transforming the time variable  $t$  into  $ct$  in the result of (17.34).

The initial value problem for the wave equation is called the **Cauchy problem** and is given by the inhomogeneous wave equation

$$\partial_t^2 u(\mathbf{x}, t) - \Delta_n u(\mathbf{x}, t) = f(\mathbf{x}, t) \quad (17.36)$$

under the two initial conditions

$$\begin{aligned}u(\mathbf{x}, 0) &= \phi(\mathbf{x}), \\ \partial_t u(\mathbf{x}, 0) &= \psi(\mathbf{x}),\end{aligned}$$

where  $f$ ,  $\phi$ ,  $\psi$  are continuous and differentiable given functions. For example,  $f(\mathbf{x}, t)$  provides an external force acting on the system described by (17.36).

The wave operator (17.33) is a linear operator, so the solution is the sum for the general solution of the homogeneous equation (17.34) and a particular solution for the inhomogeneous equation (17.36).

### 17.4.2 Homogeneous Wave Equations

First, we provide the solution for the one-dimensional homogeneous version ( $f = 0$ ) of the Cauchy problem (17.36), in which the spatial part is defined on the whole region of one dimension  $-\infty < x < \infty$ . For example, consider the case of an infinitely long vibrating string. The wave equation is written as

$$u_{tt} - u_{xx} = 0, \quad (17.37)$$

which is a hyperbolic second-order PDE that we can express by

$$\left(\frac{\partial}{\partial t} - \frac{\partial}{\partial x}\right) \left(\frac{\partial}{\partial t} + \frac{\partial}{\partial x}\right) u = 0. \quad (17.38)$$

Let us set

$$u_t + u_x = v, \quad (17.39)$$

then the first-order PDE for  $v(t, x)$  is obtained from (17.38) as

$$v_t - v_x = 0. \quad (17.40)$$

As shown earlier, (17.40) has a solution of the form

$$v(x, t) = g(x + t), \quad (17.41)$$

where  $g$  is any function. Thus we must solve (17.39) for  $u$ , which is given by

$$u_t + u_x = g(x + t). \quad (17.42)$$

One solution of (17.42) takes the following form:

$$u(x, t) = h(x + t), \quad (17.43)$$

which we can check through direct differentiation of (17.43) by setting  $p = x + t$  such that

$$\frac{\partial u}{\partial x} = \frac{dh}{dp} \frac{\partial p}{\partial x} = h',$$

$$\frac{\partial u}{\partial t} = \frac{dh}{dp} \frac{\partial p}{\partial t} = h'.$$

Then we have

$$h(p) = \frac{1}{2} \int^p g(p) dp. \quad (17.44)$$

Another possibility is the general solution of the homogeneous equation obtained by setting  $g = 0$  in (17.42), which takes the form

$$u = z(x - t). \quad (17.45)$$

Adding this to (17.44), we have the general expression of a solution,

$$u(x, t) = h(x + t) + z(x - t). \quad (17.46)$$

Now let us solve (17.46) under the initial conditions

$$\begin{aligned} u(x, 0) &= \phi(x), \\ u_t(x, 0) &= \psi(x), \end{aligned}$$

where  $\phi$  and  $\psi$  are given functions of  $x$ . From (17.46), we have the relations

$$\phi(x) = h(x) + z(x), \quad (17.47)$$

$$\psi(x) = h'(x) - z'(x). \quad (17.48)$$

By differentiating (17.47), we obtain  $\phi' = h' + z'$ . Combining this with (17.48), we have

$$h' = \frac{1}{2}(\phi' + \psi), \quad z' = \frac{1}{2}(\phi' - \psi).$$

Integrating on  $p$  yields

$$y(p) = \frac{1}{2}\phi(p) + \frac{1}{2} \int_0^p \psi dp + a,$$

$$z(p) = \frac{1}{2}\phi(p) + \frac{1}{2} \int_0^p \psi dp - a.$$

So, we get

$$u(x, t) = \frac{1}{2} [\phi(x + t) + \phi(x - t)] + \frac{1}{2} \int_{x-t}^{x+t} \psi(p) dp, \quad (17.49)$$

which is the solution for the initial value problem for the homogeneous equation (17.37).

### 17.4.3 Inhomogeneous Wave Equations

Next we solve the initial value problem for an inhomogeneous PDE ( $f \neq 0$ ) by applying the **method of characteristic coordinates**. We transform the variables  $x, t$  into new variables  $\xi = x + t$ ,  $\eta = x - t$ . The wave equation for the new variables yields

$$\partial_\eta \partial_\xi u = -\frac{1}{4} f\left(\frac{\xi + \eta}{2}, \frac{\xi - \eta}{2}\right).$$

This equation can be integrated with respect to  $\eta$ , leaving  $\xi$  as a constant. Thus we have

$$u_\xi = -\frac{1}{4} \int^\eta f d\eta, \quad (17.50)$$

where the lower limit of integration is arbitrary. Again we can integrate with respect to  $\xi$ :

$$u(\xi, \eta) = -\frac{1}{4} \int^\xi \int^\eta f\left(\frac{\xi + \mu}{2}, \frac{\xi - \eta}{2}\right) d\eta d\xi. \quad (17.51)$$

Here we consider the dependent variable  $u$  at a fixed point  $(\xi_0, \eta_0)$  defined by

$$\xi_0 = x_0 + t_0, \quad \eta_0 = x_0 - t_0. \quad (17.52)$$

We can evaluate (17.51) at the point  $(\xi_0, \mu_0)$  and make a particular choice of the lower limits such that

$$u(\xi_0, \mu_0) = \frac{1}{4} \int_{-\infty}^{\xi_0} \int_{\eta_0}^{\infty} f d\eta d\xi.$$

Here we change the variables  $\xi, \eta$  into the original ones  $(x, t)$ , and the Jacobian is the determinant of its coefficient matrix:

$$J = \det \begin{vmatrix} \frac{\partial \xi}{\partial x} & \frac{\partial \xi}{\partial t} \\ \frac{\partial \eta}{\partial x} & \frac{\partial \eta}{\partial t} \end{vmatrix} = 2.$$

Thus  $d\xi d\eta = J dx dt = 2 dx dt$ , so the double integral can be transformed as

$$u(x_0, t_0) = \frac{1}{2} \int_0^{t_0} \int_{x_0-(t_0-t)}^{x_0+(t_0-t)} f(x, t) dx dt$$

As a result, we have the following theorem:

♠ **Theorem:**

The unique solution of (17.36) on one spatial dimension is given by

$$u(x, t) = \frac{1}{2} [\phi(x + t) + \psi(x - t)] \\ + \frac{1}{2} \int_{x-t}^{x+t} \psi(p) dp + \frac{1}{2} \int_0^t \int_{x-(t-t')}^{x+(t-t')} f(x', t') dx' dt',$$

where  $\phi(x) = u(x, 0)$  and  $\psi(x) = u_t(x, 0)$ .

#### 17.4.4 Wave Equations in Finite Domains

In this section we attempt to solve the wave equations defined in the region  $D \times (0, \infty)$ , where  $D$  is the bounded domain of  $\mathbf{R}^n$ . For this problem, we have to specify the initial conditions at  $t = 0$  as well as some boundary conditions on  $\partial D$ . As noted in Sect. 17.1.2, the commonly used boundary conditions are the **Dirichlet** and **Neumann** conditions. First we treat a homogeneous wave equation with no external term given by

$$\partial_t^2 u - \Delta_n u = 0, \quad (17.53)$$

where the initial condition is

$$u(\mathbf{x}, 0) = f(\mathbf{x}), \quad \partial_t u(\mathbf{x}, 0) = g(\mathbf{x}), \quad (17.54)$$

and the boundary conditions on  $\partial D$  are given by

$$u(\mathbf{x}, t) = 0 \quad \text{or} \quad \partial_n u(\mathbf{x}, t) = 0. \quad (17.55)$$

Thus, when the boundary conditions are independent of  $t$ , the **method of separation of variables** is useful, i.e., we assume that the solution  $u$  takes the form

$$u(\mathbf{x}, t) = X(\mathbf{x})T(t), \quad (17.56)$$

where  $X$  satisfies the boundary conditions (17.55) on  $\partial D$ . Substituting (17.56) into (17.53), we have

$$-\frac{\Delta X(\mathbf{x})}{X(\mathbf{x})} = -\frac{T''(t)}{T(t)} = \mu^2. \quad (17.57)$$

This defines a quantity  $\mu^2$  that must be constant since  $\Delta X/X$  depends only on  $\mathbf{x}$  and  $T''/T$  depends only on  $t$ . The reason for the positive constant  $\mu^2 > 0$  will be seen later.

Equation (17.57) gives a pair of separate differential equations for  $X(\mathbf{x})$  and  $T(t)$ : the one equation is

$$\Delta X(\mathbf{x}) = -\mu^2 X(\mathbf{x}) \quad (17.58)$$

that satisfies the given boundary conditions of (17.55), and the other is

$$T''(t) = -\mu^2 T(t) \quad (17.59)$$

in which  $0 < t < \infty$ . The solution for this ODE is obtained as

$$T(t) = a \cos \mu t + b \sin \mu t, \quad (17.60)$$

where  $a$  and  $b$  are constants that can be determined from the initial conditions.

Combining (17.60) with  $X(\mathbf{x})$ , the solution is expressed as

$$u(\mathbf{x}, t) = X_\mu(\mathbf{x})(a \cos \mu t + b \sin \mu t). \quad (17.61)$$

This is a **normal mode** of vibration with eigenfrequency  $\mu$  and the general solution is obtained by the superposition of **normal modes**. Thus, we have the general solution of the form

$$u(\mathbf{x}, t) = \sum_n X_n(\mathbf{x})(a_n \cos \mu_n t + b_n \sin \mu_n t). \quad (17.62)$$

For example, from the initial conditions in (17.54), we have

$$\sum_n a_n X_n = f, \quad \sum_n \mu_n b_n X_n = g,$$

so the coefficients in (17.62) are given by

$$a_n = \langle f | X_n \rangle, \quad b_n = \langle g | F_n \rangle / \mu_n.$$

## Exercises

1. Find the general solution for the wave equation defined on the one-dimensional bounded domain  $(0, l) \times (0, \infty)$ , which is given by

$$\partial_t^2 u - \Delta u = 0,$$

under the conditions:  $u(x, 0) = f(x)$ ,  $\partial_t u(x, 0) = g(x)$ ,  $u(0, t) = u(l, t) = 0$ .

**Solution:** The normalized eigenfunctions are  $X_n = \sqrt{\frac{2}{l}} \sin\left(\frac{n\pi x}{l}\right)$ , and the associated eigenfrequencies  $\mu_n$  are the integer multiples of the fundamental frequency  $\pi/l$ . Thus, we obtain

$$u(x, t) = \sum \left( a_n \cos \frac{n\pi t}{l} + b_n \sin \frac{n\pi t}{l} \right) \sin \frac{n\pi x}{l},$$

where the coefficients are

$$a_n = \frac{2}{l} \int_0^l f(x) \sin \frac{n\pi x}{l} dx, \quad b_n = \frac{1}{\pi n} \int_0^l g(x) \sin \frac{\pi x}{l} dx. \quad \clubsuit$$

2. The differential equation under a point source  $\mathbf{r} = 0$  at time  $t_0 = 0$  in an infinite medium is given by

$$\partial_t G - \Delta_3 G = \delta(\mathbf{r})\delta(t). \quad (17.63)$$

Find the solution  $G(\mathbf{r}, t)$  called **Green's function** by means of Fourier and Laplace transforms.

**Solution:** If we take a Fourier transform in space and a Laplace transform in time, (17.63) becomes  $G(\mathbf{k}, \omega) = 1/[(2\pi)^3(\omega + k^2)]$ . The inverse Laplace transform yields  $G(\mathbf{k}, t) = e^{-k^2 t}/(2\pi)^3$ . We then obtain Green's function  $G(\mathbf{r}, t)$  by the spatial Fourier transform given by

$$G(\mathbf{r}, t) = \frac{1}{(2\pi)^3} \int e^{-k^2 t} e^{i\mathbf{k} \cdot \mathbf{r}} d^3 \mathbf{k}.$$

Integrating over the angles of  $\mathbf{r}$  yields

$$\begin{aligned} G(\mathbf{r}, t) &= \frac{1}{(2\pi)^2} \int_0^\infty e^{-k^2 t} \frac{\sin kr}{kr} k^2 dk = \frac{1}{4\pi r} \operatorname{Im} \int_{-\infty}^\infty e^{-k^2 t} e^{ikr} k dk \\ &= \frac{1}{4\pi r} \operatorname{Im} \frac{1}{i} \frac{\partial}{\partial r} \int_{-\infty}^\infty e^{-k^2 t} e^{ikr} dk, \end{aligned}$$

which gives Green's function in the form  $G(\mathbf{r}, t) = e^{-r^2/4t}/(4\pi t)^{3/2}$ ,  $t > 0$ . For  $t < 0$ ,  $G(\mathbf{r}, t) = 0$ . ♣

3. Find the half-space one-dimensional Green's function defined on  $x > 0$  that satisfies the boundary condition of  $G(x, t) = 0$  at  $x_0 = 0$ .

**Solution:** Using an image source of negative strength at  $x = -x_0$ , the solution is expressed by  $G(x, t) = G_0(x - x_0, t) - G_0(x + x_0, t)$ ,  $x > 0$ , where  $G_0(x, t) = e^{-x^2/4t}/(4\pi t)^{1/2}$ . ♣

4. Consider the wave equation with a source term  $h(\mathbf{r}, t)$  given by

$$\partial_t^2 f - \Delta_3 f = h(\mathbf{r}, t).$$

Show that the solution is expressed as

$$f(\mathbf{r}, t) = \frac{1}{4\pi} \int h(\mathbf{r}', t) \frac{1}{|\mathbf{r} - \mathbf{r}'|} d^3 \mathbf{r}'. \quad (17.64)$$

**Solution:** The Green function  $G(\mathbf{r}, t)$  is defined as the solution satisfying the equation  $\partial_t^2 G - \Delta_3 G = \delta(\mathbf{r})\delta(t)$ . The spatial Fourier and temporal Laplace transform of the above is obtained by settings  $\omega \rightarrow -\omega^2$  in Exercise 1 as  $G(\mathbf{k}, \omega) = 1/[(2\pi)^3(-\omega^2 + k^2)]$ . The inverse transform of the above gives  $G(|\mathbf{r} - \mathbf{r}'|, t - t') = \delta(|\mathbf{r} - \mathbf{r}'|)\delta(t - t')/(4\pi|\mathbf{r} - \mathbf{r}'|)$ . Since the physical system is invariant under the translations in space and time, the Green function depends only on relative space and time coordinates  $|\mathbf{r} - \mathbf{r}'|$  and  $t - t'$ . The Green function has the property that the solution can be written as  $f(\mathbf{r}, t) = \int G(\mathbf{r}, \mathbf{r}'; t, t')h(\mathbf{r}', t')d^3\mathbf{r}'dt'$ , so it is given by (17.64). ♣

6. Find the general solution for the wave equation

$$\partial_t^2 u - \Delta_3 u = 0,$$

where  $\Delta_3$  is the Laplacian defined by

$$\Delta_3 = \partial_x^2 + \partial_y^2 + \partial_z^2.$$

**Solution:** Since the system is isotropic and homogeneous, we can assume the solution in analogy with the one-dimensional case as  $f(p) = f(\mathbf{n} \cdot \mathbf{x} \pm t)$ , where  $\mathbf{n}$  is the unit vector that points along the direction of propagation of the wave and  $\mathbf{n} \cdot \mathbf{x} = lx + my + nz$ . From the chain rule on derivatives, we have  $(l^2 + m^2 + n^2 - 1)f''(p) = 0$ . Thus,  $f$  can be arbitrary since  $l^2 + m^2 + n^2 = 1$ . The general solution is given by  $u(\mathbf{x}, t) = f(\mathbf{n} \cdot \mathbf{x} - t) + g(\mathbf{n} \cdot \mathbf{x} + t)$ , where  $\mathbf{n}$  is any unit vector. ♣

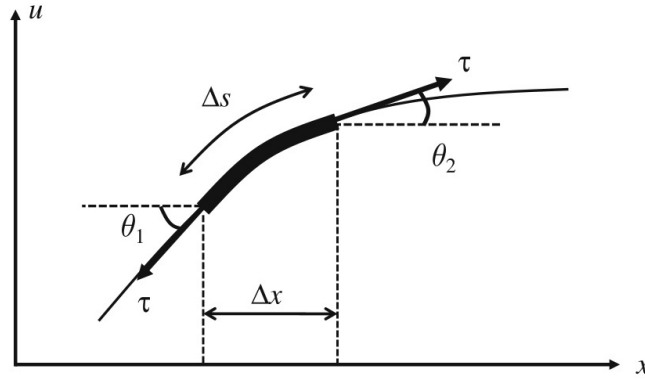
## 17.5 Applications in Physics and Engineering

### 17.5.1 Wave Equations for Vibrating Strings

In the previous sections, we rigorously studied the theories underlying the following three typical classes of partial differential equations: Laplace equations, diffusion equations, and wave equations. In this section, we attempt to formulate mathematical expressions for these classes on the basis of the associated physical phenomena; e.g., we will see that the mathematical form of the wave equation

$$\partial_t^2 u(x, t) = v^2 \partial_x^2 u(x, t) \quad (17.65)$$

is derived by considering the wavy motion of a that string. This will make clear why equations of the same form as (17.65) are called *wave* equations.



**Fig. 17.1.** Schematic of a thin stretched string and a line element  $\Delta s$ ; the tensions exerted at both ends of the line element have equal magnitudes  $\tau$  but act in different directions

Suppose that a thin string is stretched between two fixed points with a tension  $\tau$  exerted at the two end points. We assume that the string is perfectly flexible, i.e., only tensile forces can be transmitted in the tangential direction. Then, as illustrated in Fig. 17.1, the magnitude of the tension exerted in the tangential direction is the same for every part of the string. It can be seen from the figure that the vertical components of the tension  $\tau$  at the two ends of a line element with length  $\Delta s$  are  $-\tau \sin \theta_1$  and  $\tau \sin \theta_2$ , where  $\tau = |\tau|$ . Hence, the vertical component  $\tau_u$  of the external force exerted on the line element is given by

$$\tau_u = \tau (\sin \theta_2 - \sin \theta_1). \quad (17.66)$$

The sine terms are rewritten in terms of the derivative  $\partial_x u$  by using the following approximations:

$$\begin{aligned} \sin \theta_1 &= \frac{\partial u}{\partial s} = \partial_x u \frac{dx}{ds} = \partial_x u \frac{1}{\sqrt{1 + (\partial_x u)^2}} \\ &\simeq \partial_x u \left[ 1 + \frac{1}{2} (\partial_x u)^2 + \cdots \right] \simeq \partial_x u \end{aligned}$$

where the relation  $ds = \sqrt{(dx)^2 + (du)^2}$  was used, and

$$\begin{aligned} \sin \theta_2 &= \frac{\partial u(x + \Delta x)}{\partial s} = \frac{\partial u(x + \Delta x)}{\partial x} \frac{dx}{ds} \simeq \frac{\partial u(x + \Delta x)}{\partial x} \\ &= \partial_x u + \partial_x^2 u|_{x=\xi} \Delta x, \end{aligned}$$

where  $\xi$  is a constant satisfying the condition  $x \leq \xi \leq x + \Delta x$ . (The **mean value theorem** ensures the existence of such a constant  $\xi$ .) Substitution of the sine terms in (17.66) yields

$$\tau_u = \tau \left. \partial_x^2 u \right|_{x=\xi} \Delta x.$$

Since the inertial force exerted by the line element is given by

$$\rho \Delta s \partial_t^2 u,$$

where, ( $\rho$  is the line density) we obtain the equation of motion for the string in the vertical direction:

$$\rho \frac{\Delta s}{\Delta x} \partial_t^2 u = \tau \left. \partial_x^2 u \right|_{x=\xi}.$$

Taking the limit  $\Delta x \rightarrow 0$  so that  $\xi \rightarrow x$ , and then approximating  $ds/dx$  as 1, we obtain the final result:

$$\partial_t^2 u = \frac{\tau}{\rho} \partial_x^2 u. \quad (17.67)$$

It is customary to denote the positive constant as  $\tau/\rho$ ,  $v^2$ , which allows us to write (17.67) in the familiar form (17.65).

### 17.5.2 Diffusion Equations for Heat Conduction

In this subsection, we attempt to derive mathematical expressions for diffusion equations by considering the physical phenomena that occur during heat conduction, i.e., the flow of heat in a certain medium from points at a high temperature to those at lower temperatures. This process takes place in such a manner that molecules in irregular motion exchange their kinetic energy by colliding with each other.

We aim to determine the amount of heat  $\delta Q$  penetrating an arbitrarily chosen surface element  $\delta S$  inside the medium per unit time (called the **heat flux**). In order to find  $\delta Q$ , we consider another surface element  $\delta S_1$ , which has the same magnitude as  $\delta S$ , parallel to  $\delta S$  and located at an orthogonal distance  $\Delta n$  from  $\delta S$ . We assume that  $\delta S$  is so small that the temperatures  $u = u(x, y, z, t)$  on  $\delta S$  and  $u_1 = u(x + \delta x, y + \delta y, z + \delta z, t)$  on  $\delta S_1$  are constant over  $\delta S$  and  $\delta S_1$ , respectively.

From thermodynamics, we know that the magnitude of the flux of heat difference between  $u$  and  $u_1$ , denoted by  $\delta u$ , and the area of the surface element are related in the following manner:

$$\delta Q \cdot \delta n = \kappa \cdot \delta u \cdot \delta S, \quad (17.68)$$

where the value of the constant  $\kappa$ , called the **thermal conductivity**, depends on the medium. Dividing both sides of (17.68) by  $\delta n$  and taking the limit  $\delta n \rightarrow 0$ , we have

$$\delta Q = \kappa \frac{\partial u}{\partial n} \cdot \delta S.$$

Here,  $\partial u / \partial n$  is the derivative in the direction normal to  $\delta S$  and is expressed as

$$\frac{\partial u}{\partial n} = \bar{\mathbf{n}} \cdot \nabla u,$$

where  $\bar{\mathbf{n}}$  is the unit vector normal to  $\delta S$ . Thus, we obtain the flux passing through a volume element  $\delta v$  that is enclosed by a surface  $S$ :

$$\begin{aligned} Q &= \kappa \iint_S \frac{\partial u}{\partial n} dS = \kappa \iint_S \nabla u \cdot \bar{\mathbf{n}} dS \\ &= \kappa \iiint_V \nabla \cdot (\nabla u) dV. \end{aligned}$$

We now apply the mean value theorem to the volume integral over  $\delta v$  to obtain

$$Q = \kappa \nabla \cdot [\nabla u(x^*, y^*, z^*, t)] \delta v, \quad (17.69)$$

where  $(x^*, y^*, z^*)$  is a point in  $\delta v$ .

Apart from the above-mentioned discussion, we also see that  $Q$  is related to the temperature variation in  $\delta v$  with time. In fact, the temperature  $u$  in  $\delta v$  increases (or decreases) owing to the accumulation (or loss) of heat in  $\delta v$  at a rate of  $\partial u / \partial t$ . Therefore, the flow of heat into (or out of)  $\delta v$  can be written as

$$\rho \sigma \frac{\partial u}{\partial t} \delta v, \quad (17.70)$$

where  $\rho$  is the mass density and  $\sigma$  (the **specific heat**) is a characteristic of the medium. By setting (17.69) equal to (17.70), and allowing the volume element  $\delta v$  to shrink to a point, we have

$$\rho \sigma \partial_t u = \kappa \nabla \cdot (\nabla u) = \kappa \nabla^2 u.$$

Clearly, this result is of the same form as a diffusion equation that describes heat conduction in a medium with physical parameters  $\rho, \sigma, \kappa$ .

**Tensor Analyses**



## Cartesian Tensors

**Abstract** Tensors are geometric entities that provide concise mathematical frameworks for formulating problems in physics and engineering. The most important feature of tensors is their coordinate invariance: tensors are independent of the type of coordinate system chosen. This feature is similar to the condition that the length and direction of a geometric figure do not change, regardless of the coordinate system used for the algebraic expression. In contrast, the components of a tensor are coordinate-dependent in a structured routine. In this chapter, we discuss the ways in which the choice of a coordinate system affects the components of a tensor.

### 18.1 Rotation of Coordinate Axes

#### 18.1.1 Tensors and Coordinate Transformations

A **tensor** is a natural generalization of a vector or a scalar encountered in elementary vector calculus. The latter two are, in fact, both special cases of tensors of order  $n$ , whose specification in a three-dimensional space requires  $3^n$  numbers, called the **components** of the tensor. In this context, scalars are tensors of the zero order with a  $3^0 = 1$  component and vectors are tensors of the first order with  $3^1 = 3$  components.

Of importance is the fact that a tensor of order  $n$  is much more than just a set of  $3^n$  numbers. The key property of tensors is adherence to the transformation law of its components under a change of coordinate system, say, from a rectangular to elliptic, polar, or other curvilinear coordinate system. If the coordinate system is changed to a new one, the components of a tensor change according to a characteristic transformation law. We shall see that this transformation law makes clear the physical (or geometrical) meaning of the tensor being invariant under a change of coordinate system. The coordinate-invariance-character of tensors answers the demand that the proper formulation of physical laws should be independent of the choice of coordinate systems.

It is obvious that physical processes must be independent of the coordinate system. However, what is not so trivial is what the coordinate-independence property of physical processes implies about the transformation law of mathematical objects (i.e., tensors). The study of these implications and the classification of physical quantities by means of the transformation laws constitute the primary content of this chapter. Emphasis is placed on the fact that all kinds of tensors are geometric objects whose representation (i.e., The values of its components) obey a characteristic transformation law under **coordinate transformation**.

### 18.1.2 Summation Convention

In order to simplify subsequent notation, we introduce the following two conventions:

♠ **Summation convention:**

When the same index appears repeatedly in one term, we carry out a summation with respect to the index. The range of summation is from 1 to  $n$ , where  $n$  is the number of dimensions of the space.

♠ **Range convention:**

All non repeated indices are understood to run from 1 to  $n$ .

These conventions are operative throughout this chapter unless specifically stated otherwise.

*Example* The summation convention yields the new notation as

$$a_i b_i \equiv \sum_{i=1}^n a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n.$$

Similarly, if  $i$  and  $j$  have the range from 1 to 2, then

$$\begin{aligned} a_{ij} b_{ij} &= a_{1j} b_{1j} + a_{2j} b_{2j} \\ &= a_{11} b_{11} + a_{12} b_{12} + a_{21} b_{21} + a_{22} b_{22}, \end{aligned}$$

where it does not matter whether the first sum is carried out on  $i$  or  $j$ .

**Remark.** Repeated indices are referred to as **dummy indices** since, owing to the implied summation, any such pair may be replaced by any other pair of repeated indices without changing the meaning of the mathematical expression.

### 18.1.3 Cartesian Coordinate System

A tensor is a mathematical object composed of several components. The values of the components depend on the coordinate system to be employed, so are altered through a coordinate transformation even when the tensor itself remains unchanged. Among the many possible choices of coordinate transformations, a **rigid rotation** of a **rectangular Cartesian coordinate system** is the simplest. The remainder of this section is devoted to explaining the basic properties of the simplest coordinate transformations, as a preliminary for our subsequent study of tensors in terms of more general coordinate systems.

We begin with two formal definitions:

♠ **Cartesian coordinate system:**

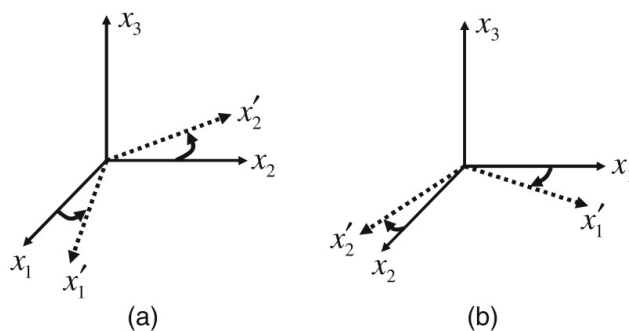
A Cartesian coordinate system associates a unique ordered set of real numbers (**coordinates**)  $(x_1, x_2, \dots, x_n)$  with every point in a given  $n$ -dimensional space by reference to a set of directed straight lines (**coordinate axes**)  $Ox_1, Ox_2, \dots, Ox_n$  intersecting at the origin  $O$ .

♠ **Rectangular Cartesian coordinate system:**

If the three axes of a Cartesian coordinate system are mutually perpendicular, we have what is called a **rectangular Cartesian coordinate system**.

Figure 18.1 is a schematic illustration of a rectangular Cartesian coordinate system in three-dimensional space. Referring to this coordinate system, we denote the triples  $(1, 0, 0)$ ,  $(0, 1, 0)$ ,  $(0, 0, 1)$  by  $e_1$ ,  $e_2$ ,  $e_3$ , respectively. These triples are represented geometrically by mutually perpendicular unit arrows.

The set of Cartesian axes  $Ox_1$ ,  $Ox_2$ , and  $Ox_3$  is said to be **right-handed** if and only if the rotation needed to turn the  $x^1$ -axis into the direction of the  $x^2$ -axis through an angle  $\angle x^1 Ox^2 < \pi$  would propel a right-handed screw toward the positive direction of the  $x^3$ -axis. Conversely, if such a rotation



**Fig. 18.1.** (a) Right-handed and (b) left-handed Cartesian coordinate systems

propels a left-handed screw in the positive direction of the  $x^3$ -axis, the set of axes is said to be **left-handed**. In this section, we consider only Cartesian coordinate systems that are rectangular and right-handed.

#### 18.1.4 Rotation of Coordinate Axes

We now formulate a rigid rotation of rectangular Cartesian axes. Assume a position arrow  $\mathbf{r}$  whose components are given by  $(x_1, x_2, x_3)$  and  $(x'_1, x'_2, x'_3)$  in terms of two different rectangular coordinate systems having a common origin. We denote the set of unit orthogonal basis arrows associated with the unprimed and primed system by  $\{\mathbf{e}_i\}$  and  $\{\mathbf{e}'_i\}$ , respectively. The transformation from one Cartesian coordinate system to another is called a **rigid rotation of Cartesian axes** and has the following property:

♠ **A rigid rotation of Cartesian axes:**

A rigid rotation of Cartesian axes is described by the transformation equations of coordinates  $x_k$  as

$$x'_j = R_{jk} x_k \quad (\text{summed over } k), \quad (18.1)$$

$$x_k = R_{jk} x'_j \quad (\text{summed over } j), \quad (18.2)$$

where  $R_{jk} = \mathbf{e}'_j \cdot \mathbf{e}_k$  are **directed cosines** of  $\mathbf{e}'_j$  associated with  $\mathbf{e}_k$ .

*Remarks.*

1. Each of the two indices  $j$  and  $k$  for  $R_{jk}$  refers to a different basis: the first index  $j$  refers to the *primed* set  $\{\mathbf{e}'_j\}$ , while the second index  $k$  refers to the *unprimed* one  $\{\mathbf{e}_k\}$ . This means that, in general,  $R_{jk} \neq R_{kj}$ .
2. The transformation coefficients  $R_{jk}$  do not constitute a tensor, but simply set of real numbers. (See the second remark in Sect. 18.2.3.)

*Proof* A geometric arrow  $\mathbf{r}$  joining the origin  $O$  and the point  $P$  is expressed by

$$\mathbf{r} = x_k \mathbf{e}_k = x'_j \mathbf{e}'_j. \quad (18.3)$$

We expand  $\mathbf{e}_k$  by the set of  $\{\mathbf{e}'_j\}$  as

$$\mathbf{e}_k = (\mathbf{e}'_j \cdot \mathbf{e}_k) \mathbf{e}'_j = R_{jk} \mathbf{e}'_j, \quad (18.4)$$

where we use both the range and the summation conventions. Substituting (18.4) into (18.3), we obtain

$$x_k R_{jk} \mathbf{e}'_j = x'_j \mathbf{e}'_j,$$

and thus

$$(x_k R_{jk} - x'_j) \mathbf{e}'_j = \mathbf{0}.$$

Since the arrow set  $\{\mathbf{e}'_j\}$  is linearly independent, the quantities in the parentheses equal zero, which results in the desired equation (18.1).

Similarly, expanding  $\{\mathbf{e}'_j\}$  by  $\{\mathbf{e}_k\}$  as

$$\mathbf{e}'_j = (\mathbf{e}_k \cdot \mathbf{e}'_j) \mathbf{e}_k = R_{jk} \mathbf{e}_k \quad (18.5)$$

and substituting it into (18.3), we arrive at equation (18.2). ♣

**Remark.** Observe that in the transformation law (18.1) and the expansion (18.5),  $R_{jk}$  acts on an *unprimed* entity (i.e.,  $x_k$  or  $\mathbf{e}_k$ ) to produce a primed one (i.e.,  $x'_j$  or  $\mathbf{e}'_j$ ). However, in (18.2) and (18.4),  $R_{jk}$  acts on a primed entity to produce an unprimed one. In all the cases above, we should make sure that the order of indices  $j$  and  $k$  attached to the coefficients  $R_{jk}$  remains unchanged: the first index,  $j$ , always refers to the primed entity and the second,  $k$ , to the unprimed one.

### 18.1.5 Orthogonal Relations

The following theorem states an important property of the transformation coefficients  $R_{jk}$  that gives rise to a rigid rotation of Cartesian axes.

#### ♠ Orthogonal relations:

The transformation coefficients  $R_{jk}$  for a rigid rotation of axes satisfy the conditions

$$R_{ik} R_{jk} = \delta_{ij}, \quad R_{ik} R_{i\ell} = \delta_{k\ell}. \quad (18.6)$$

*Proof* The first relation of (18.6) follows from a geometric formula for the angle  $\theta$  between two basic arrows:  $\mathbf{e}'_i$  and  $\mathbf{e}'_j$ . Taking the inner product of the two basic arrows  $\mathbf{e}'_i = R_{ik} \mathbf{e}_k$  and  $\mathbf{e}'_j = R_{j\ell} \mathbf{e}_\ell$ , we have

$$\cos \theta = \mathbf{e}'_i \cdot \mathbf{e}'_j = R_{ik} R_{j\ell} (\mathbf{e}_k \cdot \mathbf{e}_\ell) = R_{ik} R_{j\ell} \delta_{k\ell} = R_{ik} R_{jk}. \quad (18.7)$$

If  $i = j$ ,  $\mathbf{e}'_i$  and  $\mathbf{e}'_j$  coincide so that  $\theta = 0$ , whereas if  $i \neq j$ ,  $\mathbf{e}'_i$  and  $\mathbf{e}'_j$  are orthogonal so that  $\theta = \pi/2$ . Hence, we have

$$R_{ik} R_{jk} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

The second equation of (18.6) can be verified in a similar manner by considering the angle between  $\mathbf{e}_k$  and  $\mathbf{e}_\ell$ . ♣

The physical meaning of the relations (18.6) is rather obvious. They ensure that the axes of each set  $\{\mathbf{e}'_i\}$  or  $\{\mathbf{e}_k\}$  are mutually orthogonal, i.e.,

$$\mathbf{e}'_i \cdot \mathbf{e}'_j = \delta_{ij} \quad \text{and} \quad \mathbf{e}_k \cdot \mathbf{e}_\ell = \delta_{k\ell}.$$

### 18.1.6 Matrix Representations

Since the transformation coefficients  $R_{jk}$  have two subscripts, it is natural to display their values in matrix form. The notation  $[R_{jk}]$  is used to denote the matrix having  $R_{jk}$  as the element in the  $j$ th row and  $k$ th column. In addition, when denoted by  $\mathbf{R}$ , it represents a linear operator of a rotation of axes without reference to any of the values of its coefficients  $R_{jk}$ .

*Example* In two dimensions, a rigid rotation of rectangular axes for an angle  $\theta$  is given by

$$\mathbf{e}'_i = R_{ij}\mathbf{e}_j, \quad [R_{ij}] = \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix} = \begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix}. \quad (18.8)$$

This rotation of axes gives rise to a coordinate transformation,

$$x'_i = R_{ij}x_j,$$

or equivalently,

$$\begin{pmatrix} x'_1 \\ x'_2 \end{pmatrix} = \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}. \quad (18.9)$$

*Remark.* We comment briefly on the distinction between **active** and **passive transformations**; since it often causes confusion. Throughout this chapter, we are concerned solely with passive transformations, for which physical entities of interest (e.g., the mass of a particle or a geometric arrow) remain unaltered and only the coordinate system is changed from  $\{\mathbf{e}_i\}$  to  $\{\mathbf{e}'_i\}$ , as given by (18.8). In contrast, an active transformation alters the position and/or the direction of the physical entity itself, while the axes  $\{\mathbf{e}_i\}$  remain fixed. In the latter case, a rotation of a geometric arrow  $\mathbf{x}$  through an angle  $\theta$  in two dimensions is described by

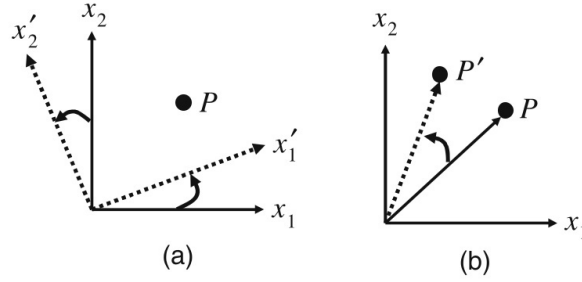
$$\mathbf{e}'_i = \mathbf{e}_i \quad \text{and} \quad \begin{pmatrix} x'_1 \\ x'_2 \end{pmatrix} = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix},$$

which obviously differs from those for a passive transformation. Figure 18.2 illustrates the difference between the two transformations.

It can be shown that the determinant of the matrix  $[R_{k\ell}]$  reads

$$\det[R_{k\ell}] = \pm 1 \quad (18.10)$$

(see the proof in Exercise 5). This means that there are two classes of rectangular Cartesian coordinate systems, corresponding to the positive and negative signs in (18.10). Throughout this section, we consider only cases of



**Fig. 18.2.** Difference between a passive (a) and an active (b) transformation

$\det[R_{k\ell}] = +1$ , which corresponds to our previous restriction to a single type of coordinate system (i.e., right-handed). We shall see in sect. 18.3.1 that the rotation of coordinate axes whose transformation coefficients give rise to  $\det[R_{k\ell}] = -1$  yields the left-handed system, which for the moment is beyond our scope.

### 18.1.7 Determinant of a Matrix

We close this section by commenting on a formal definition of the determinant of a matrix and its relevant materials, as a preliminary for the exercises below.

#### 1. The determinant

$$D = \det[a_{ij}] = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix} \quad (18.11)$$

of the square array of  $n^2$  numbers (elements)  $a_{ij}$  is the sum of the  $n!$  terms

$$(-1)^r a_{1k_1} a_{2k_2} \cdots a_{nk_n}, \quad (18.12)$$

each corresponding to one of the  $n!$  different ordered sets  $k_1, k_2, \dots, k_n$  obtained by  $r$  interchanges of elements from the ordered set  $1, 2, \dots, n$ .

2. The **minor** (or **complementary minor**)  $M_{ij}$  of the elements  $a_{ij}$  in the  $n$ th-order determinant  $D = \det[a_{ij}]$  is the  $(n - 1)$ th-order determinant obtained from (18.11) on erasing the  $i$ th row and the  $j$ th column.

*Example* Given a third-order determinant

$$D = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix},$$

its  $a_{12}$  minor is obtained by removing the first row and the second column in  $D$ , as expressed by

$$M_{12} = \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix}.$$

3. The **cofactor**  $C_{ij}$  of the element  $a_{ij}$  is defined by

$$C_{ij} = \frac{\partial D}{\partial a_{ij}},$$

or equivalently,

$$C_{ij} = (-1)^{i+j} M_{ij}.$$

4. A determinant  $D$  may be represented in terms of the elements and cofactors of any one row or column by

$$D = \det[a_{ij}] = \sum_{i=1}^n a_{ij} C_{ij} = \sum_{k=1}^n a_{jk} C_{jk} \quad (\text{with } j \text{ fixed}). \quad (18.13)$$

This is called **simple Laplace development** of a determinant  $D$ . (The proof is given in Exercise 2). The expression (18.13) gives the same value for  $D$  regardless of the column or row [i.e., no matter what value of  $j$  in (18.13)] we choose in the expansion. Note also that for  $j \neq h$ ,

$$\sum_{i=1}^n a_{ij} C_{ih} = \sum_{k=1}^n a_{jk} C_{hk} = 0.$$

*Example* The expansion by the first row gives

$$D = \begin{vmatrix} 1 & 3 & 0 \\ 2 & 6 & 4 \\ -1 & 0 & 2 \end{vmatrix} = 1 \begin{vmatrix} 6 & 4 \\ 0 & 2 \end{vmatrix} - 3 \begin{vmatrix} 2 & 4 \\ -1 & 2 \end{vmatrix} + 0 \begin{vmatrix} 2 & 6 \\ -1 & 0 \end{vmatrix} = -12.$$

**Remark.** In view of the expansion (18.13), an  $n$ th-order determinant  $D$  is represented by a linear combination of  $n$  numbers of  $(n-1)$ th-order determinants. Similarly, each of the latter  $(n-1)$ th-order determinants is in turn represented in terms of  $n-1$  numbers of  $(n-2)$ th-order determinants, and so on. In a successive manner, we finally arrive at  $n!$  numbers of first-order determinants (i.e., just  $n!$  real numbers), each of which is expressed by (18.12).

## Exercises

1. Check the validity of the orthogonal conditions (18.6) for the transformation coefficients  $R_{ij}$  in two dimensions.

**Solution:** For instance, if we set  $j = 1$  and  $k = 2$ , then

$$R_{i1}R_{i2} = R_{11}R_{12} + R_{21}R_{22} = -\cos \theta \sin \theta + \cos \theta \sin \theta = 0,$$

or if  $j = k = 2$ , we have

$$R_{i2}R_{i2} = R_{21}R_{21} + R_{22}R_{22} = (-\sin \theta)^2 + \cos^2 \theta = 1.$$

Other equations can be proved in a similar way. ♣

2. Show that the expression (18.13) for a determinant

$$D = \det[a_{pq}] = \sum_{k=1}^n a_{ik} C_{ik} \quad \text{and} \quad D = \sum_{i=1}^n a_{\ell i} C_{\ell i} \quad \text{for fixed } i$$

yields the same value of  $D$  no matter what value of  $i$  we choose.

**Solution:** We prove only the first formula, since the proof of the second is quite similar to that of the first.

It easily follows that the statement is true for a second-order determinant for which the expansions with fixed  $i = 1$ ,  $a_{11}a_{22} + a_{12}(-a_{21})$  and that with  $i = 2$ ,  $a_{21}(-a_{12}) + a_{22}a_{11}$ , give the same value. By mathematical induction, we tentatively assume that the statement is true for an  $(n-1)$ th-order determinant and try to prove that it is also true for an  $n$ th-order determinant.

To do so, we expand  $D$  in terms of each of two arbitrary rows, say, the  $i$ th and the  $j$ th row with  $i < j$ , and compare the results.

(i) Let us first expand  $D$  by its  $i$ th row. A typical term in the expansion by the  $i$ th row reads

$$a_{ik} C_{ik} = a_{ik} \cdot (-1)^{i+k} M_{ik}, \quad (18.14)$$

where  $i$  is fixed and  $k$  runs from 1 to  $n$ . Since the minor  $M_{ik}$  of  $a_{ik}$  in  $D$  is an  $(n-1)$ th-order determinant, owing to the induction hypothesis it can be expanded by any row. Expand  $M_{ik}$  by its  $(j-1)$ th row. This row corresponds to the  $j$ th row of  $D$ , because  $M_{ik}$  does not contain elements of the  $i$ th row of  $D$  and  $i < j$ . Hence, the expansion of  $M_{ik}$  by its  $(j-1)$ th row consists of the linear combination of the elements  $a_{j\ell}$  with  $\ell = 1, 2, \dots, k-1, k+1, \dots, n$  (i.e.,  $\ell \neq k$ ). We distinguish between the two cases,  $\ell < k$  and  $\ell > k$ , as follows.

For  $\ell < k$ , the element  $a_{j\ell}$  belongs to the  $\ell$ th column of  $M_{ik}$ . Hence, the term involving  $a_{j\ell}$  in the expansion of  $M_{ik}$  reads

$$a_{j\ell} \cdot (\text{cofactor of } a_{j\ell} \text{ in } M_{ik}) = a_{j\ell} \cdot (-1)^{(j-1)+\ell} M_{ikj\ell}. \quad (18.15)$$

Here  $M_{ikj\ell}$  is the minor of  $a_{j\ell}$  in  $M_{ik}$ , which is obtained from  $D$  by deleting the  $i$ th and  $j$ th rows and the  $k$ th and  $\ell$ th columns of  $D$ . Then it follows from (18.14) and (18.15) that the resulting terms in the expansion of  $D$  are of the form

$$a_{ik} a_{j\ell} \cdot (-1)^b M_{ikj\ell} \quad \text{with} \quad b = i + k + j + \ell - 1. \quad (18.16)$$

If  $\ell > k$ , the only difference is that  $a_{j\ell}$  belongs to the  $(\ell-1)$ th column of  $M_{ik}$  because  $M_{ik}$  does not contain elements of the  $k$ th column of  $D$  and  $k < \ell$ . This results in an additional minus sign

in (18.15) and instead of (18.16) we obtain  $-a_{ik}a_{j\ell} \cdot (-1)^{b+1}M_{ikj\ell}$  with the same value of  $b$ . In short, the expansion of  $D$  by the  $i$ th row yields

$$D = \sum_{k=1}^n a_{ik}C_{ik} = \sum_{k=1}^n a_{ik} \left[ \sum_{\ell=1}^{k-1} (-1)^b a_{j\ell} M_{ikj\ell} + \sum_{\ell=k+1}^n (-1)^{b+1} a_{j\ell} M_{ikj\ell} \right] \quad (18.17)$$

with  $b = i + k + j + \ell - 1$ .

(ii) We next expand  $D$  by the  $j$ th row. A typical term in this expansion is

$$a_{j\ell}C_{j\ell} = a_{j\ell} \cdot (-1)^{j+\ell} M_{j\ell}. \quad (18.18)$$

By the induction hypothesis, we may expand the minor  $M_{j\ell}$  of  $a_{j\ell}$  in  $D$  by its  $i$ th row, which corresponds to the  $i$ th row of  $D$  since  $j > i$ .

For  $k > \ell$ , the element  $a_{ik}$  in that row belongs to the  $(k-1)$ th column of  $M_{j\ell}$ , because  $M_{j\ell}$  does not contain elements of the  $\ell$ th column of  $D$  and  $\ell < k$ . Hence, the term involving  $a_{ik}$  in this expansion is

$$a_{ik} \cdot (\text{cofactor of } a_{ik} \text{ in } M_{j\ell}) = a_{ik} \cdot (-1)^{i+(k-1)} M_{ikj\ell}, \quad (18.19)$$

where the minor  $M_{ikj\ell}$  of  $a_{ik}$  in  $M_{j\ell}$  is obtained by deleting the  $i$ th and  $j$ th rows and the  $k$ th and  $\ell$ th columns of  $D$ , and is thus identical to  $M_{ikj\ell}$  in (18.15). It follows from (18.18) and (18.19) that this yields a representation whose terms are identical to those given by (18.16) when  $\ell < k$ .

For  $k < \ell$ , the element  $a_{ik}$  belongs to the  $k$ th column of  $M_{j\ell}$ , so we get an additional minus sign and the result agrees with that characterized by (18.16). Hence, we conclude that the expansion of  $D$  by the  $j(> i)$ th row,  $\sum_{k=1}^n a_{jk}C_{jk}$ , is identical to the expansion (18.17).

The conclusions from the discussions in (i) and (ii) clearly show that the two expansions of  $D$  consist of the same terms, which completes our proof of the given statement. ♣

3. Let  $b_{kj} = C_{jk}/D$ , where  $C_{jk}$  is the cofactor of  $[a_{jk}]$  in  $D = \det[a_{jk}]$ . Show that

$$b_{kj}a_{\ell k} = \delta_{j\ell} \quad \text{and} \quad b_{kj}a_{j\ell} = \delta_{k\ell}, \quad (18.20)$$

which means that the matrix  $[b_{kj}]$  is the **inverse** of  $[a_{jk}]$ .

**Solution:** It follows that

$$b_{kj}a_{\ell k} = \frac{C_{jk}a_{\ell k}}{D} = \frac{a_{\ell 1}C_{j1} + a_{\ell 2}C_{j2} + \cdots + a_{\ell n}C_{jn}}{D}. \quad (18.21)$$

The discussion in Exercise 2 tells us that the sum in the numerator equals  $D$  when  $\ell = j$  regardless of the value of  $j$ . Hence, we have

$$b_{kj}a_{\ell k} = 1 \quad \text{if } j = \ell.$$

We next consider the case of  $j \neq \ell$ . To do thus we replace the elements in the  $j$ th row of  $D$  by those in the  $\ell(\neq j)$ th row of  $D$ . The resulting determinant, denoted by  $\tilde{D}$ , reads

$$\tilde{D} = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \vdots & \vdots & & \vdots \\ a_{j-1,1} & a_{j-1,2} & \cdots & a_{j-1,n} \\ a_{\ell,1} & a_{\ell,2} & \cdots & a_{\ell,n} \\ a_{j+1,1} & a_{j+1,2} & \cdots & a_{j+1,n} \\ \vdots & \vdots & & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n} \end{vmatrix} = 0,$$

since  $\tilde{D}$  has two identical rows. Note that the expansion of  $\tilde{D}$  by its  $\ell$ th row is  $\tilde{D} = \sum_{p=1}^n a_{\ell p} C_{jp}$ , which equals the sum in the numerator in (18.21). It thus follows that

$$b_{kj}a_{\ell k} = 0 \quad \text{if } j \neq \ell.$$

These arguments complete the proof of the first equation. The second can be verified in the same manner. ♣

4. Suppose a matrix  $[Q_{kj}]$  defined by  $Q_{kj} = C_{jk}/\det[R_{jk}]$ , where the  $R_{jk}$  are the transformation coefficients of a rigid rotation of axes and the  $C_{jk}$  are the cofactors of  $[R_{jk}]$  in  $\det[R_{jk}]$ . Prove that  $Q_{kj} = R_{jk}$ .

**Solution:** Apply the result of Exercise 3 to find that  $Q_{kj}R_{\ell k} = \delta_{\ell j}$ . Multiplying both sides by  $R_{\ell m}$  and summing with respect to  $\ell$ , we arrive at

$$Q_{kj}R_{\ell k}R_{\ell m} = \delta_{\ell j}R_{\ell m} = R_{jm}.$$

Then, the orthogonal relation  $R_{\ell k}R_{\ell m} = \delta_{km}$  implies that  $Q_{kj}\delta_{km} = Q_{mj} = R_{jm}$ . ♣

5. Show that  $\det[R_{jk}] = \pm 1$ .

**Solution:** It follows from (18.20) that

$$\det[Q_{kj}R_{j\ell}] = \det[\delta_{k\ell}] = 1,$$

where the  $Q_{kj}$  are the same quantities as in Exercise 4. From elementary linear algebra, we find that

$$\det[Q_{kj}R_{j\ell}] = \det[Q_{kj}]\det[R_{j\ell}] = (\det[R_{j\ell}])^2,$$

where the identity  $Q_{kj} = R_{jk}$  was used to obtain the last term. Combining the two results above, we obtain  $\det[R_{jk}] = \pm 1$ . ♣

## 18.2 Cartesian Tensors

### 18.2.1 Cartesian Vectors

Having dealt with the rotation of coordinate axes, we are ready to introduce the concept of tensors and their transformation law in terms of Cartesian coordinate systems. Assume an ordered set of three quantities  $v_i$  ( $i = 1, 2, 3$ ) that are explicit or implicit functions of  $x_j$ . Let us see how the values of  $v_i(x_j)$  change through a rigid rotation of the Cartesian axes. If they transform according to the law given below, the quantities  $v_i$  are called the **components** of a particular kind of tensor, i.e., of a **Cartesian vector** (or a **first-order Cartesian tensor**).

#### ♠ Cartesian vectors:

A Cartesian vector  $\mathbf{v}$  is an object represented by an ordered set of  $n$  functions  $v_i(x_j)$  in terms of the  $x$ -coordinate system and by another set of  $n$  functions  $v'_i(x'_j)$  in terms of the  $x'$ -coordinate system, where  $v'_i$  and  $v_i$  at each point are related by the transformation law:

$$v'_i = R_{ij}v_j \quad \text{and} \quad v_i = R_{ki}v'_k, \quad (18.22)$$

where  $R_{ij} = \mathbf{e}'_i \cdot \mathbf{e}_j$ .

Obviously, a vector  $\mathbf{v}$  is a geometric object (like an arrow) so that it is uniquely determined independently of the coordinate system. On the other hand, the function form of the components  $v_i(x_j)$  depend on our choice of coordinate system, even when we consider only the same vector  $\mathbf{v}$ . This is why the concepts of a *vector* and its *components* are inherently different from each other. (See also Sect. 18.2.2 on this point.)

We emphasize again that the index  $i$  of  $R_{ij}$  in (18.22) refers to the dashed (transformed) function  $v'_i$ , whereas the  $j$  refers to the undashed (original) one  $v_j$ . In the following, we consider several examples of ordered sets of functions  $v_i$  in two dimensions, which may or may not be a first-order Cartesian tensor.

**Examples 1.** The ordered set of functions  $(v_i)$  ( $i = 1, 2$ ) with the components  $v_1 = x_2$  and  $v_2 = -x_1$ .

Using the transformation law of coordinates  $x'_i = R_{ij}x_j$ , we set the following for each function:

$$\begin{aligned}v'_1 &= x'_2 = R_{21}x_1 + R_{22}x_2 = -x_1 \sin \theta + x_2 \cos \theta, \\v'_2 &= -x'_1 = -R_{11}x_1 - R_{12}x_2 = -x_1 \cos \theta - x_2 \sin \theta.\end{aligned}$$

On the other hand, the functions  $v'_i$  should be obtained from  $v_i$  through the transformation law as

$$\begin{aligned}v'_1 &= R_{1k}v_k = v_1 \cos \theta + v_2 \sin \theta = x_2 \cos \theta - x_1 \sin \theta, \\v'_2 &= R_{2k}v_k = -v_1 \sin \theta + v_2 \cos \theta = -x_2 \sin \theta - x_1 \cos \theta.\end{aligned}$$

The two expressions for  $v'_1$  and  $v'_2$  are identical to one another regardless of the values of  $\theta$ . Therefore, the pair of functions  $v_i(x_j)$  are components of a Cartesian vector.

2. The set  $v_i$  with  $v_1 = x_2$  and  $v_2 = x_1$ .

Following the same procedure as above, we have

$$\begin{aligned}v'_1 &= x'_2 = -sx_1 + cx_2, \\v'_2 &= x'_1 = cx_1 + sx_2\end{aligned}$$

and

$$\begin{aligned}v'_1 &= cv_1 + sv_2 = cx_2 + sx_1, \\v'_2 &= -sv_1 + cv_2 = -sx_2 + cx_1,\end{aligned}$$

where  $c$  and  $s$  represent  $\cos \theta$  and  $\sin \theta$ , respectively. These two sets of expressions do not agree with each other. Hence, the pair  $(x_2, x_1)$  is not a first-order Cartesian tensor.

### 18.2.2 A Vector and a Geometric Arrow

The result of Example 2 in Sect. 18.2.1 might be confusing for some readers; the functions  $v_i(x_1, x_2)$  given there are not components of a vector, although they *appear* to represent a geometric arrow in  $(x_1, x_2)$ -plane. To make this point clear, we have to comment on the difference between the formal definition of a *vector* as a first-order tensor and our familiar definition of a vector as a *geometric arrow*.

In elementary calculus, vectors are simply defined by a geometric arrow with certain length and direction, commonly denoted by a bold-face letter, say,  $\mathbf{v}$ . Owing to this definition,  $\mathbf{v}$  is uniquely determined by specifying its length and direction, which are both independent of our choice of coordinate systems. However, the uniqueness disappears if it is defined algebraically by specifying its components  $v_k$  relative to given coordinate axes. The values of the components  $v_k$  depend on our choice of coordinate system even when the same arrow  $\mathbf{v}$  is considered. Hence, when we apply a coordinate transformation, the values of  $v_k$  are altered in a way that preserves the length and direction of the arrow  $\mathbf{v}$ , according to (18.22), which is why we call the set of  $n$  functions  $v_k$  not a *vector*, but the *components* of a vector.

In short, we should always keep in mind that a vector is a geometric object independent of coordinate systems, whereas components of a vector are just mathematical representations of the vector with reference to a specific

coordinate system. This caution applies to all the classes of tensors presented throughout this section.

**Remark.** Despite the above caution, we sometimes call a set of components of a tensor just a “tensor” to shorten our sentences. However, it is important to note an inherent difference between a tensor (=coordinate-*independent* object) and components of a tensor (=coordinate-*dependent* quantities).

### 18.2.3 Cartesian Tensors

We turn to a **second-order Cartesian tensor** that requires two subscripts to identify a particular element of the set.

#### ♠ Second-order Cartesian tensor:

A second-order Cartesian tensor  $\mathbf{T}$  is an object represented by an ordered set of two-index quantities  $T_{ij}$  in terms of the  $x$ -coordinate system and by another set of quantities  $T'_{kl}$  in terms of the  $x'$ -coordinate system, where  $T_{ij}$  and  $T'_{kl}$  at each point are related by

$$T'_{ij} = R_{ik}R_{jl}T_{kl}, \quad T_{kl} = R_{mk}R_{nl}T'_{mn}. \quad (18.23)$$

Here, the two-index quantities  $T_{ij}$  are called **components of the tensor**  $\mathbf{T}$ .

In a similar way, we may define a Cartesian tensor of general order as follows: The set of expressions  $T_{ij\dots k}$  are the components of a Cartesian tensor if, for all rotations of the axes, the expressions using the new coordinates  $T'_{\ell m\dots n}$  are given by

$$T'_{ij\dots k} = R_{il}R_{jm}\cdots R_{kn}T_{\ell m\dots n}$$

and

$$T_{\ell m\dots n} = R_{pl}R_{qm}\cdots R_{rn}T'_{pq\dots r}.$$

It is apparent that an  $n$ th-order Cartesian tensor in three dimensions has  $3^n$  components.

*Example* Assume two Cartesian vectors  $\mathbf{a}$  and  $\mathbf{b}$ , each of which is represented by the components  $a_j$  and  $b_k$  associated with the same coordinate system. Then, it is possible to create nine products of the components expressed by

$$a_j b_k \quad (j, k = 1, 2, 3),$$

which is called an **outer product** (or **direct product**) of the vectors  $\mathbf{a}$  and  $\mathbf{b}$  (see also Sect. 18.4.3). The outer product consists of a second-order Cartesian tensor. In fact, since each  $a_i$  and  $b_j$  transforms as

$$a'_i = R_{ik}a_k \quad \text{and} \quad b'_j = R_{j\ell}b_\ell,$$

we have

$$T'_{ij} \equiv a'_ib'_j = R_{ik}R_{j\ell}a_kb_\ell = R_{ik}R_{j\ell}T_{k\ell}.$$

**Remark.** We emphasize that transformation coefficients, say, the  $R_{ij}$ , do not form a tensor and note the fact that the two indices  $i$  and  $j$  in the tensor  $T_{ij}$  refer to the *same* coordinate system, whereas those in the coefficients  $R_{ij}$  refer to *different* coordinate systems. Hence,  $T_{ij}$  and  $R_{ij}$  are inherently different from each other, though both require two indices.

### 18.2.4 Scalars

Contrary to the case of finite-order tensors, we now consider quantities that are unchanged by a rotation of axes, which are called **scalars** or **tensors of zero order** and contain only one component. The most obvious example is the square of the distance of a point from the origin, which must be invariant under any rotation of coordinate axes. Other examples of scalars are presented below.

**Examples 1.** We show that the **scalar product**  $\mathbf{u} \cdot \mathbf{v}$  is invariant under rotation. In the original (unprimed) system, the scalar product is given in terms of components by  $u_iv_i$  and in the rotated (primed) system, it is given by

$$u'_iv'_i = (R_{ij}u_j)(R_{ik}v_k) = R_{ij}R_{ik}u_jv_k = u_jv_k\delta_{jk} = u_jv_j, \quad (18.24)$$

where the **orthogonal relation**  $R_{ij}R_{ik} = \delta_{jk}$  in (18.6) was used. Since the expression in the rotated system is the same as that in the original system, the scalar product is indeed invariant under rotations.

- 2.** If the  $v_i$  are the components of a vector, the divergence  $\nabla \cdot \mathbf{v} = \partial v_i / \partial x_i$  becomes a scalar. This is proven as follows: In the rotated coordinate system,  $\nabla \cdot \mathbf{v}$  is given by

$$\frac{\partial v'_i}{\partial x'_i} = \frac{\partial}{\partial x'_i} (R_{ik}v_k) = R_{ik} \frac{\partial v_k}{\partial x'_i},$$

where the elements  $R_{ik} = \mathbf{e}_k \cdot \mathbf{e}'_i$  are not functions of position. Using the relation  $\partial x_j / \partial x'_i = R_{ij}$  (see Exercise 2 below), we have

$$R_{ik} \frac{\partial v_k}{\partial x'_i} = R_{ik} \frac{\partial x_j}{\partial x'_i} \frac{\partial v_k}{\partial x_j} = R_{ik}R_{ij} \frac{\partial v_k}{\partial x_j} = \delta_{jk} \frac{\partial v_k}{\partial x_j} = \frac{\partial v_j}{\partial x_j}.$$

Finally, we obtain

$$\frac{\partial v'_i}{\partial x'_i} = \frac{\partial v_j}{\partial x_j}.$$

## Exercises

1. Examine whether or not the ordered set of functions  $v_i$  defined by  $v_1 = (x_1)^2$  and  $v_2 = (x_2)^2$  constitute a vector. Here  $(x_1)^2$  means the square of  $x_1$ .

**Solution:** To examine the first function,  $v_1$ , alone is sufficient to show that this pair is not a vector. Evaluating  $v'_1$  directly gives

$$v'_1 = (x'_1)^2 = c^2(x_1)^2 + 2c(-s)x_1x_2 + (-s)^2(x_2)^2,$$

whereas (18.8) requires that  $v'_1 = cv_1 - sv_2 = c(x_1)^2 - s(x_2)^2$ , which is different from the above. ♣

2. Show that  $R_{ij} = \mathbf{e}'_i \cdot \mathbf{e}_j = \frac{\partial x'_i}{\partial x_j} = \frac{\partial x_j}{\partial x'_i}$  in Cartesian coordinate systems.

**Solution:** Set  $x'_i \mathbf{e}'_i = x_j \mathbf{e}_j$  and differentiate both sides with respect to  $x'_j$  to obtain  $(\partial x'_i / \partial x_j) \mathbf{e}'_i = \mathbf{e}_j$ . Taking the scalar product of both sides with  $\mathbf{e}'_k$  yields

$$\frac{\partial x'_i}{\partial x_j} \mathbf{e}'_i \cdot \mathbf{e}'_k = \frac{\partial x'_i}{\partial x_j} \delta_{ik} = \frac{\partial x'_k}{\partial x_j} \quad \text{and} \quad \mathbf{e}_j \cdot \mathbf{e}'_k = R_{kj}.$$

Hence, we have  $R_{kj} = \partial x'_k / \partial x_j$ . Similarly, if we differentiate by  $x'_i$  (instead of  $x_j$ ) the first identity yields  $R_{kj} = \partial x_j / \partial x'_k$ . ♣

3. Show that the **gradient of a vector**  $\mathbf{v}$ , denoted by  $\nabla \mathbf{v}$ , is a second-order tensor.

**Solution:** Suppose that  $v_i$  represents the components of a vector  $\mathbf{v}$  and consider the quantities generated by  $T_{ij} = \partial v_i / \partial x_j$  ( $i, j = 1, 2, 3$ ). These nine quantities form the components of a second-order tensor, as can be seen from the fact that

$$T'_{ij} = \frac{\partial v'_i}{\partial x'_j} = \frac{\partial (R_{ik} v_k)}{\partial x'_j} \frac{\partial x_\ell}{\partial x'_j} = R_{ik} \frac{\partial v_k}{\partial x_\ell} R_{j\ell} = R_{ik} R_{j\ell} T_{k\ell}. \quad \clubsuit$$

**Remark.** The concept (and its notation)  $\nabla \mathbf{v}$  introduced above is not in the category of simple vector calculus. In fact, the quantity  $\nabla \mathbf{v}$  is not a vector like  $\nabla \times \mathbf{v}$  and  $\nabla \phi$ , but a second-order tensor.

## 18.3 Pseudotensors

### 18.3.1 Improper Rotations

So far our coordinate transformations have been restricted to rigid rotations described by an orthogonal matrix  $[R_{ij}]$  with the property

$$|\mathbf{R}| \equiv \det[R_{ij}] = +1.$$

Such transformations are called **proper rotations**. We now broaden our discussion to include transformations that are described by an orthogonal matrix  $[R_{ij}]$  for which

$$|\mathbf{R}| = -1.$$

The latter kind of transformations are called **improper rotations** (or **rotation with reflection**). Below are two examples of improper rotations.

### (a) Inversion

The most obvious example of an improper rotation is an **inversion** of the coordinate axes through the origin represented by

$$\mathbf{e}'_i = -\mathbf{e}_i \quad \text{for all } i = 1, 2, 3.$$

In this case, a position arrow  $\mathbf{x}$  is described in terms of the bases  $\mathbf{e}'_i$  and  $\mathbf{e}_i$ , respectively, by

$$\mathbf{x} = x_i \mathbf{e}_i \quad \text{and} \quad \mathbf{x} = x'_i \mathbf{e}'_i = x'_i (-\mathbf{e}_i).$$

Equating them, we obtain

$$x_i = -x'_i = -\delta_{ij} x'_j,$$

which shows that an inversion of axes is expressed by  $R_{ij} = -\delta_{ij}$ . In fact, its determinant becomes

$$|\mathbf{R}| = \begin{vmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{vmatrix} = -1.$$

### (b) Reflection

Another example is a **reflection** that reverses the direction of one basis:

$$\mathbf{e}'_i = -\mathbf{e}_i \quad \text{for a specified } i.$$

For the reflection of the  $x$ -axis, e.g., we have

$$|\mathbf{R}| = \begin{vmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix} = -1.$$

**Remark.** Note that an inversion is different from a proper rotation only for an odd dimensionality. In a case of two or four dimensions, for instance, an inversion is the same as a proper rotation. In contrast, a reflection that changes the sign of only one coordinate is always different from a proper rotation regardless of the dimensionality.

Through an improper rotation, our initial **right-handed coordinate system** is changed into a **left-handed** one. This is illustrated schematically in Fig. 18.3. The reader should note that such a change cannot be accomplished by any kind of proper rotation.

### 18.3.2 Pseudovectors

Regardless of whether it is proper or improper, any rotation described by  $R_{ij}$  transforms the components  $v_i$  of a vector  $\mathbf{v}$  as

$$v'_i = R_{ij}v_j.$$

This is because any real physical vector  $\mathbf{v}$  may be considered as a geometrical object (i.e., an arrow in space), whose direction and magnitude cannot be altered merely by describing it in terms of a different coordinate system.

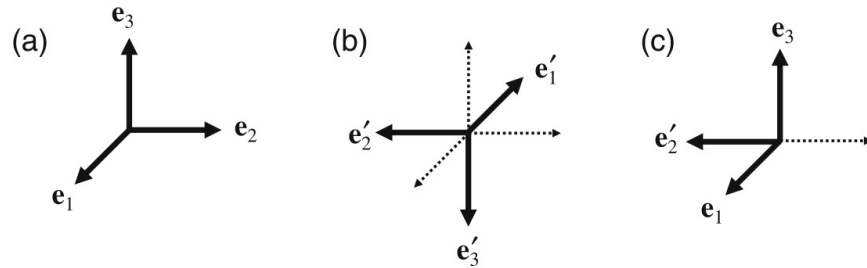
It is, however, possible to define another type of vector  $\mathbf{w}$  whose components  $w_i$  transform as

$$\begin{aligned} w'_i &= R_{ij}w_j && \text{under proper rotations,} \\ w'_i &= -R_{ij}w_j && \text{under improper rotations,} \end{aligned}$$

or equivalently,

$$w'_i = |R|R_{ij}w_j.$$

In this case, the  $w_i$  are no longer strictly the components of a true Cartesian vector. Rather, they are said to form the components of a **pseudovector** or a first-order Cartesian **pseudotensor**. A pseudovector may be alternatively referred to as an **axial vector**; correspondingly, a true vector may be called a **polar vector**.



**Fig. 18.3.** Improper rotation: (b) inversion and (c) reflection of the right-handed Cartesian coordinate system depicted in (a)

**Remark.** A pseudovector should not be considered a real *geometric arrow* in space, since its direction is reversed by an improper transformation of the coordinate axes. This is illustrated in Fig. 18.4, where the pseudovector  $\mathbf{w}$  is shown as a broken line to indicate that it is *not* a real physical vector.

Below is a summary of the discussion above.

♠ **Vectors and pseudovectors:**

Components  $v_i$  of a vector  $\mathbf{v}$  transform as

$$v'_i = R_{ij}v_j$$

under a rigid rotation of Cartesian axes, whereas components  $w_i$  of a pseudovector  $\mathbf{w}$  transform as

$$w'_i = |\mathbf{R}|R_{ij}w_j,$$

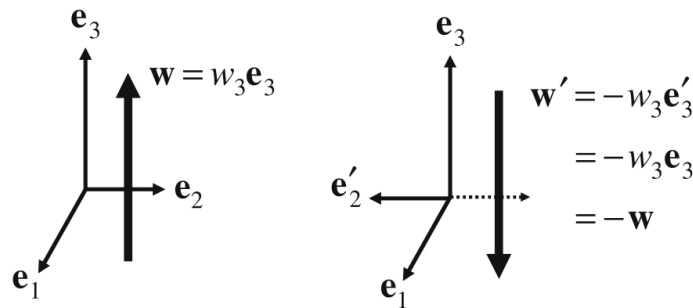
where  $|\mathbf{R}|$  is the determinant of the transformation matrix  $[R_{ij}]$ .

Hence, the difference between a vector and a pseudovector manifests when applying an improper rotation that yields  $|\mathbf{R}| = -1$ .

Pseudovectors occur frequently in physics, although this fact is not usually pointed out explicitly. Following are physical examples of pseudovectors.

*Examples* The following three physical quantities are all pseudovectors.

1. Angular momentum of a moving particle,  $\mathbf{L} = \mathbf{r} \times \mathbf{p}$ , where  $\mathbf{r}$  is the particle's position arrow and  $\mathbf{p}$  its moment vector.
2. Torque on a particle,  $\mathbf{N} = \mathbf{r} \times \mathbf{F}$ , where  $\mathbf{r}$  is the particle's position arrow and  $\mathbf{F}$  the force acting on the particle.
3. Magnetic field,  $\mathbf{B} = \nabla \times \mathbf{A}$ , defined by the rotation of the vector potential  $\mathbf{A}$ .



**Fig. 18.4.** Reversing behavior of the pseudovector  $\mathbf{w}$  via the reflection of the  $\mathbf{e}_2$ -axis

It is noteworthy that each of these pseudovectors consists of a vector product of two vectors.

### 18.3.3 Pseudotensors

We can extend the notion of vectors and pseudovectors to objects with two or more subscripts. For instance, assume a quantity with components transforming as

$$T'_{ij} = R_{ik}R_{j\ell}T_{k\ell}$$

under proper rotations, but

$$T'_{ij} = -R_{ik}R_{j\ell}T_{k\ell}$$

under improper ones. Then, the  $T_{ij}$  are components of a **second-order Cartesian pseudotensor**. Similarly, Cartesian pseudotensors of arbitrary order are defined such that their components transform as

$$T'_{ij\dots k} = |\mathbf{R}|R_{i\ell}R_{jm}\cdots R_{kn}T_{\ell m\dots n}$$

where  $|\mathbf{R}|$  is the determinant of the transformation matrix  $[R_{ij}]$ . Corresponding to these, zeroth-order objects may also be divided into scalars and pseudoscalars, the latter being invariant under rotation but changing sign on improper rotation.

### 18.3.4 Levi–Civita Symbols

A typical example of a third-order pseudotensor is the **Levi–Civita symbol**  $\varepsilon_{ijk}$ .

#### ♠ Levi–Civita symbol:

The Levi–Civita symbol (or the **permutation symbol**), denoted by  $\varepsilon_{ijk}$ , takes the values  $+1$  and  $-1$  if the ordered set  $i, j, k$  is obtained by an even or odd permutation, respectively, of the set  $1, 2, 3$ .

Actually,  $\varepsilon_{ijk}$  takes the values

$$\begin{aligned}\varepsilon_{123} &= \varepsilon_{231} = \varepsilon_{312} = +1, \\ \varepsilon_{213} &= \varepsilon_{321} = \varepsilon_{132} = -1,\end{aligned}$$

and  $\varepsilon_{ijk} = 0$  if any two of the indices  $i, j, k$  are equal.

The pseudotensor property of  $\varepsilon_{ijk}$  follows from a convenient notation for the determinant  $|\mathbf{A}|$  of a general  $3 \times 3$  matrix  $[A_{ij}]$  (see Exercise 2):

$$|\mathbf{A}|\varepsilon_{lmn} = A_{i\ell}A_{jm}A_{kn}\varepsilon_{ijk}.$$

Certainly, this equation holds for the transformation matrix  $[R_{ji}]$  for rigid rotation. Hence, we have

$$|R|\varepsilon_{lmn} = R_{il}R_{jm}R_{kn}\varepsilon_{ijk},$$

or equivalently,

$$\varepsilon_{ijk} = |R|R_{il}R_{jm}R_{kn}\varepsilon_{lmn}. \quad (18.25)$$

This shows that  $\varepsilon_{ijk}$  is a third-order Cartesian tensor.

The result (18.25) indicates more than the pseudotensorial character of  $\varepsilon_{ijk}$ . It clearly demonstrates that all of the components of  $\varepsilon_{ijk}$  are unaltered by any rotation of axes. Tensors endowed with this property are called **isotropic tensors** (**invariant tensors** or **fundamental tensors**). We know that there are no isotropic tensors of first order and that the only ones of second and third order are scalar multiples of  $\delta_{ij}$  and  $\varepsilon_{ijk}$ , respectively. Additionally, the most general isotropic tensor of fourth order is given by

$$\lambda\delta_{ik}\delta_{mp} + \mu\delta_{im}\delta_{kp} + \nu\delta_{ip}\delta_{km},$$

with arbitrary constants  $\lambda, \mu, \nu$ . (Such a fourth-order isotropic tensor occurs in the **elasticity theory** of solids; see Sect. 18.5.4). All the isotropic tensors above are relevant to the description of the physical properties of an isotropic medium (i.e., a medium having the same properties regardless of the way in which it is orientated).

## Exercises

1. Show that an angular momentum  $\mathbf{L} = \mathbf{r} \times \mathbf{p}$  is a pseudovector.

**Solution:** Since the position vector  $\mathbf{r}$  and the momentum vector  $\mathbf{p}$  are vectors, they transform under certain rotations of the axes (proper and improper) as  $r'_j = R_{jk}r_k$ ,  $p'_m = R_{mn}p_n$ . Hence, the components of  $\mathbf{L}$  in a new coordinate system read

$$\begin{aligned} L'_i &= \varepsilon'_{ijk}r'_jp'_k \\ &= (|R|R_{il}R_{jm}R_{kn}\varepsilon_{lmn})(R_{jq}r_q)(R_{ks}p_s) \\ &= |R|R_{il}(R_{jm}R_{jq})(R_{kn}R_{ks})\varepsilon_{lmn}r_qp_s \\ &= |R|R_{il}\delta_{mq}\delta_{ns}\varepsilon_{lmn}r_qp_q \\ &= |R|R_{il}\varepsilon_{lmn}r_m p_n = |R|R_{il}L_\ell, \end{aligned}$$

which clearly indicates that the quantities  $L_i$  form the components of a first-order Cartesian pseudotensor (i.e., a pseudovector). ♣

2. Determine whether  $|A|\varepsilon_{lmn} = A_{il}A_{jm}A_{kn}\varepsilon_{ijk}$  holds for a general matrix  $[A_{ij}]$  in three dimensions.

**Solution:** Set  $\ell = 1$ ,  $m = 2$ ,  $n = 3$ , for instance, to find that the right-hand side reads

$$\begin{aligned} A_{i1}A_{j2}A_{k3}\varepsilon_{ijk} &= A_{11}A_{22}A_{33} + A_{21}A_{32}A_{13} + A_{31}A_{12}A_{23} \\ &\quad - A_{11}A_{32}A_{23} - A_{21}A_{12}A_{33} - A_{31}A_{22}A_{13} = |A|. \end{aligned}$$

Other cases can be proven in the same manner. ♣

3. Derive the identity:  $\varepsilon_{ijk}\varepsilon_{klm} = \delta_{il}\delta_{jm} - \delta_{im}\delta_{jl}$ .

**Solution:** We first note that the right-hand side of the above identity,  $\delta_{il}\delta_{jm} - \delta_{im}\delta_{jl}$ , reads

$$+1 \quad \text{if } i = l \text{ and } j = m \neq i, \quad (18.26)$$

$$-1 \quad \text{if } i = m \text{ and } j = l \neq i, \quad (18.27)$$

$$0 \quad \text{otherwise.}$$

In the case of (18.26), the left-hand side of the desired identity is

$$\varepsilon_{ijk}\varepsilon_{klm} = \varepsilon_{ijk}\varepsilon_{kij} = (\varepsilon_{ijk})^2. \quad (18.28)$$

Since  $i \neq j$ , (18.28) takes the value +1 when  $k \neq i$  and  $k \neq j$ . As a result, we successfully obtain the desired identity. A similar procedure reveals that  $\varepsilon_{ijk}\varepsilon_{klm} = -1$  in the case of (18.27) and 0 otherwise. ♣

**Remark.** We should note that in (18.28), we have not summed with respect to  $i$  and  $j$ . This is because the second term in (18.28) was obtained by a substitution of particular values into the subscripts  $l$  and  $m$ , respectively.

## 18.4 Tensor Algebra

### 18.4.1 Addition and Subtraction

We demonstrate below the bases of tensor algebra that provide ways of constructing new tensors from old ones. For convenience, we may simply refer to  $T_{ij}$  as the tensor, but it should always be remembered that the  $T_{ij}$  are the components of  $\mathbf{T}$  in a specific coordinate system.

The addition and subtraction of tensors are defined in an obvious fashion. If  $A_{ij\dots k}$  and  $B_{ij\dots k}$  are (the components of) tensors of the same order, then their sum and differences,  $S_{ij\dots k}$  and  $D_{ij\dots k}$  respectively, are given by

$$S_{ij\dots k} = A_{ij\dots k} + B_{ij\dots k},$$

$$D_{ij\dots k} = A_{ij\dots k} - B_{ij\dots k},$$

for each set of values  $i, j, \dots k$ . Furthermore, the linearity of a rotation of coordinates immediately yields

$$\begin{aligned} R_{ip}S_{pq\dots r} &= R_{ip}(A_{pq\dots r} + B_{pq\dots r}) = R_{ip}A_{pq\dots r} + R_{ip}B_{pq\dots r} \\ &= A'_{iq\dots r} + B'_{iq\dots r} = S'_{iq\dots r}. \end{aligned}$$

### 18.4.2 Contraction

Next is an operation peculiar to tensor algebra that is of considerable importance in certain manipulations.

#### ♠ Contraction:

Contraction is an operation that makes two of the indices equal and sums over all values of the equalized indices.

As an example, we consider a third-order tensor  $T_{ijk}$  whose transformation law is described by

$$T'_{ijk} = R_{il}R_{jm}R_{kn}T_{lmn}. \quad (18.29)$$

Now we perform a contraction of this tensor with respect to  $j$  and  $k$ . Setting  $j = k$  in (18.29) and summing over  $k$ , we get

$$T'_{ikk} = R_{il}R_{km}R_{kn}T_{lmn} = R_{il}\delta_{mn}T_{lmn} = R_{il}T_{lnn},$$

where we used the orthogonality condition on the sum  $R_{km}R_{kn}$ . The result indicates that the quantity  $T_{ikk}$  forms the components of a tensor of order  $1 = 3 - 2$ . In general, contraction reduces the order of a tensor by two; contraction of an  $N$ th-order tensor  $T_{ij\dots l\dots m\dots k}$  by making the subscripts  $l$  and  $m$  equal produces another tensor of order  $N - 2$ . In particular, if contraction is applied to a tensor of order 2, the result is a scalar.

### 18.4.3 Outer and Inner Products

Let us consider the multiplication of tensors. For example, we may take two tensors  $A_{ij}$  and  $B_{k\ell m}$  of different order and simply write them in juxtaposition:

$$C_{ijk\ell m} \equiv A_{ij}B_{k\ell m}. \quad (18.30)$$

Then, the quantities are the components of a tensor of fifth-order, which follows immediately from the transformation law of tensors. Such a product

of (18.30), in which all the indices are different from one another, is called an **outer product** of tensors.

Another kind of tensor, product of known as the **inner product** of tensors, is obtained from the outer product by contraction. For instance, putting  $j = k$  in (18.30) results in

$$C_{ijj\ell m} \equiv A_{ij}B_{j\ell m}, \quad (18.31)$$

which consists of a third-order tensor as demonstrated in Sect. 18.4.2. Then, the right-hand side of (18.31) is called an inner product of the components of the tensors  $A_{ij}$  and  $B_{k\ell m}$ .

*Examples* The process of taking the scalar product of two vectors  $\mathbf{u}$  and  $\mathbf{v}$ , expressed by  $u_i v_i$ , can be recast into tensor language as forming the outer product

$$T_{ij} \equiv u_i v_j$$

and then contracting it to give

$$T_{ii} = u_i v_i.$$

Using the concept of outer (and inner) product of tensors, we can write many familiar expressions of vector algebra as contracted tensors. For example, the vector product  $\mathbf{a} = \mathbf{b} \times \mathbf{c}$  has

$$a_i = \varepsilon_{ijk} b_j c_k,$$

as its  $i$ th component, where  $\varepsilon_{ijk}$  is the Levi-Civita symbol introduced in Sect. 18.3.4. This notation clarifies the distinction between the pseudovector consisting of the components  $\varepsilon_{ijk} b_j c_k$  and the second-order tensor composed of the outer product  $b_i c_j$ .

*Remark.* The outer product of two vectors is often denoted without reference to any coordinate system as

$$\mathbf{T} = \mathbf{u} \otimes \mathbf{v}. \quad (18.32)$$

This should not be confused with the vector product of two vectors, which is itself a pseudovector and is discussed in Sect. 18.3.2. The expression (18.32) gives the basis to which the components  $T_{ij}$  of the second-order tensor refer: since  $\mathbf{u} = u_i \mathbf{e}_i$  and  $\mathbf{v} = v_j \mathbf{e}_j$ , we may write the tensor  $\mathbf{T}$  as

$$\mathbf{T} = u_i \mathbf{e}_i \otimes v_j \mathbf{e}_j = u_i v_j \mathbf{u} \otimes \mathbf{v} = T_{ij} \mathbf{u} \otimes \mathbf{v}.$$

Furthermore, we have

$$\mathbf{T} = u_i \mathbf{e}_i \otimes v_j \mathbf{e}_j = u'_i \mathbf{e}'_i \otimes v'_j \mathbf{e}'_j,$$

which indicates that the quantities  $T'_{ij}$  are the components of the same tensor  $\mathbf{T}$  but referred to a different coordinate system. These concepts can be extended to higher-order tensors.

We show below several expressions of vector algebra as contracted Cartesian tensors: the notation  $[\mathbf{a}]_i$  indicates that one takes the  $i$ th component of the vector (or tensor)  $\mathbf{a}$ .

*Examples*

1.  $\mathbf{a} \cdot \mathbf{b} = a_i b_i = \delta_{ij} a_i b_j.$
2.  $[\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})]_i = \delta_{il} a_i [\mathbf{b} \times \mathbf{c}]_l = \delta_{il} a_i (\varepsilon_{ljk} b_j c_k) = \varepsilon_{ijk} a_i b_j c_k.$
3.  $\nabla^2 \phi = \frac{\partial^2 \phi}{\partial x_i \partial x_i} = \delta_{ij} \frac{\partial^2 \phi}{\partial x_i \partial x_j}.$
4.  $[\nabla \times \mathbf{v}]_i = \varepsilon_{ijk} \frac{\partial v_k}{\partial x_j}.$
5.  $[\nabla(\nabla \cdot \mathbf{v})]_i = \frac{\partial}{\partial x_i} (\nabla \cdot \mathbf{v}) = \frac{\partial}{\partial x_i} \left( \frac{\partial v_j}{\partial x_j} \right) = \delta_{jk} \frac{\partial^2 v_j}{\partial x_i \partial x_k}.$
6.  $[\nabla \times (\nabla \times \mathbf{v})]_i = \varepsilon_{ijk} \frac{\partial}{\partial x_j} [\nabla \times \mathbf{v}]_k = \varepsilon_{ijk} \varepsilon_{klm} \frac{\partial^2 v_m}{\partial x_j \partial x_l}.$

#### 18.4.4 Symmetric and Antisymmetric Tensors

The order of subscripts attached to a tensor is important; in general,  $T_{ij}$  is not the same as  $T_{ji}$ . But there are some cases of interest as described below.

##### ♠ Symmetric and asymmetric tensor:

If

$$T_{ij} = T_{ji},$$

holds for all  $i$  and  $j$ , the tensor composed of  $T_{ij}$  is called a **symmetric tensor**. Otherwise, if

$$T_{ij} = -T_{ji}, \quad (18.33)$$

the tensor is said to be **antisymmetric** (or **skew-symmetric**).

A tensor that is symmetric (or antisymmetric) in one coordinate system remains symmetric (or antisymmetric) in any other coordinate system. In fact, if  $T_{ij}$  is symmetric in a given system, i.e.,  $T_{ij} = T_{ji}$ , then

$$T'_{ij} = R_{ik} R_{j\ell} T_{k\ell} = R_{j\ell} R_{ik} T_{\ell k} = T_{ji},$$

and similarly for antisymmetry and tensors of higher order.

Notably, every tensor can be resolved into symmetric and antisymmetric parts by the identity

$$T_{ij} = S_{ij} + A_{ij}, \quad (18.34)$$

where

$$S_{ij} = \frac{1}{2}(T_{ij} + T_{ji}) \quad \text{and} \quad A_{ij} = \frac{1}{2}(T_{ij} - T_{ji}).$$

Evidently  $S_{ij}$  is a symmetric tensor since it is unaltered even if  $i$  and  $j$  are interchanged. In contrast,  $A_{ij}$  is an antisymmetric tensor since the signs of all the components are reversed by exchanging  $i$  and  $j$ . Then,  $S_{ij}$  and  $A_{ij}$  are called the **symmetric** and **antisymmetric parts** of  $T_{ij}$ , respectively.

#### 18.4.5 Equivalence of an Antisymmetric Second-Order Tensor to a Pseudovector

It is noteworthy that in three dimensions, a second-order **antisymmetric tensor**  $\mathbf{W}$  is associated with a pseudovector  $\mathbf{w}$ . To see this, let the  $W_{ij}$  be components of an antisymmetric second-order tensor whose the transformation law reads

$$\begin{aligned} W'_{ij} &= R_{il}R_{jm}W_{\ell m} \\ &= R_{i1}R_{j2}W_{12} + R_{i1}R_{j3}W_{13} + R_{i2}R_{j1}W_{21} + R_{i2}R_{j3}W_{23} \\ &\quad + R_{i3}R_{j1}W_{31} + R_{i3}R_{j2}W_{32}, \end{aligned} \quad (18.35)$$

since  $W_{11} = W_{22} = W_{33} = 0$ . Moreover, since  $W_{\ell m} = -W_{m\ell}$ , we can reduce (18.35) to the form

$$W'_{ij} = \sum_{(\ell, m)} (R_{il}R_{jm} - R_{im}R_{j\ell})W_{\ell m}, \quad (18.36)$$

where the sum  $\sum_{(\ell, m)}$  restricts the values of  $(\ell, m)$  to  $(1, 2)$ ,  $(2, 3)$ , or  $(3, 1)$ .

Now we introduce the notation

$$w_1 \equiv W_{23} = W_{32}, \quad w_2 \equiv W_{31} = W_{13}, \quad w_3 \equiv W_{12} = W_{21},$$

or more concisely,

$$w_n \equiv W_{\ell m},$$

where  $\ell, m, n$  is a cyclic permutation of the numbers 1, 2, 3, i.e.,

$$(\ell, m, n) = (1, 2, 3), (2, 3, 1), (3, 1, 2).$$

Then (18.36) can be written as

$$w'_k = \sum_{(\ell, m, n)} (R_{il}R_{jm} - R_{im}R_{j\ell})w_n, \quad (18.37)$$

in which  $i, j, k$  and  $\ell, m, n$  are both cyclic permutations of 1, 2, 3.

Noteworthy is the fact that (18.37) is equivalent to the transformation law of components  $w_k$  of a pseudovector  $\mathbf{w}$ . After some algebra, we see that equation (18.37) can be reduced to a more compact form as

$$w'_k = |R|R_{kn}w_n, \quad (18.38)$$

which is nothing but the transformation law of a pseudovector. [See Exercise 2 for the proof of (18.38).]

We have now arrived at the following theorem:

♠ **Theorem:**

Assume a second-order antisymmetric tensor in three dimensions, whose components  $W_{ij}$  take the form

$$[W_{ij}] = \begin{pmatrix} 0 & W_{12} & -W_{31} \\ -W_{12} & 0 & W_{23} \\ W_{31} & -W_{23} & 0 \end{pmatrix}.$$

Then, the three components,  $W_{12}$ ,  $W_{31}$ , and  $W_{23}$  can be associated with the pseudovector  $\mathbf{w}$  whose components are given by

$$(w_1, w_2, w_3) = (W_{23}, W_{31}, W_{12}),$$

or more concisely,

$$w_i = \frac{1}{2}\varepsilon_{ijk}W_{jk}. \quad (18.39)$$

The right-hand side of (18.39) is a twice-contracted product of the third-order pseudotensor,  $\varepsilon_{ijk}$ , and second-order tensor,  $W_{ij}$ ; hence, it is a pseudovector.

*Examples* In physical applications, we often use the vector representation (18.39) of a second-order antisymmetric tensor. For instance, let us consider the equations of angular momentum of a moving particle with mass  $m$ . We assume that a force  $\mathbf{F}$  acts on the particle located at  $\mathbf{x}$ . Then, with  $i$  and  $j$  each taking the values 1, 2, 3 we get

$$m(\ddot{x}_j x_k - \ddot{x}_k x_j) = F_j x_k - F_k x_j, \quad (18.40)$$

which gives us nine equations. Note that both sides of (18.40) are antisymmetric tensors. Among the nine equations, therefore, there are only three that are independent,  $(j, k) = (1, 2), (2, 3), (3, 1)$ . So we can convert (18.40) into a more concise vector form as

$$mw_i = N_i,$$

where we have defined

$$w_i = \varepsilon_{ijk}(\ddot{x}_j x_k - \ddot{x}_k x_j) \quad \text{and} \quad N_i = \varepsilon_{ijk}(f_j x_k - f_k x_j).$$

### 18.4.6 Quotient Theorem

Sometimes it is necessary to clarify whether a set of functions, say,  $\{a_i(x_j)\}$ , forms the components of a vector or not. A direct method is to examine whether the functions satisfy a required transformation law under a rotation of axes, which is, however, troublesome in practice. In this subsection, we describe an alternative and more efficient method, called the **quotient law**, which is a simple indirect test for determining whether a given set of quantities forms the components of a tensor.

♠ **Quotient theorem:**

If  $a_i v_i$  is a scalar for a vector  $\mathbf{v}$  in any rotated coordinate system, then the  $a_i$  constitute the components of a vector  $\mathbf{a}$ .

*Proof* Suppose that we are given a set of  $n$  quantities  $a_i$  subject to the condition that  $a_i v_i$  is a scalar for components  $v_i$  of arbitrary vector  $\mathbf{v}$  in terms of an *arbitrarily rotated* coordinate system. We may then write

$$a_j v_j = \phi, \quad (18.41)$$

in which  $\phi$  denotes a scalar. Denoting the (as yet unknown) transform of  $a_i$  by  $a'_j$ , we know that in the  $x'$ -coordinate system the condition (18.41) reads

$$a'_i v'_i = \phi'. \quad (18.42)$$

Since  $\phi$  is a scalar,  $\phi = \phi'$ . Furthermore, since  $v_i$  are components of a vector, it follows that

$$v'_i = R_{ij} v_j.$$

Accordingly, subtracting (18.42) from (18.41) gives

$$(a_j - a'_i R_{ij}) v_j = 0. \quad (18.43)$$

On the left-hand side, a summation over  $j$  is implied, so, we cannot assert directly that the coefficients of  $v_j$  vanish. However, since (18.43) should be valid for any coordinate system, we may specifically choose the coordinate system in which the components of  $\mathbf{v}$  read  $v_1 = 1$  and  $v_{(i \neq 1)} = 0$ . Equation (18.43) then reduces to

$$a_1 - R_{i1} a'_i = 0.$$

Similarly, choosing an appropriately rotated coordinate system that provides the components  $v_2 = 1$  and  $v_{(i \neq 2)} = 0$ , we infer that

$$a_2 - R_{i2} a'_i = 0.$$

Continuing in this manner, we find that

$$a_j = R_{ij}a'_i \quad \text{for all } j.$$

Multiplying both sides by  $R_{kj}$  yields

$$R_{kj}a_j = R_{kj}R_{ij}a'_i = \delta_{ki}a'_i = a'_k,$$

i.e.,

$$a'_i = R_{ij}a_j,$$

which is the transformation law for components of a vector. We thus conclude that the  $a_i$  constitute the components of a vector, denoted by  $\mathbf{a}$ . ♣

**Remark.** In applications of the above theorem, one must be certain that the coordinate system employed is arbitrarily rotated, and this hypothesis represents a very strict condition that is not often satisfied.

#### 18.4.7 Quotient Theorem for Two-Subscripted Quantities

As a second important case, assume a set of  $n^2$  quantities  $a_{ij}$  such that  $a_{ij}v_iv_j$  is a scalar  $\phi$  for a vector  $\mathbf{v}$  and for any rotated coordinate system. Our task is to examine whether such two-subscripted quantities  $a_{ij}$  constitute the components of a tensor of second order. We shall see, however, that the answer is negative. In fact, we can say nothing about the tensorial character of  $a_{ij}$  from the hypothesis noted above, which implies the need to modify the quotient theorem for two-subscripted quantities.

Developing the modified quotient theorem requires a discussion that parallels that given in Sect. 18.4.5. By hypothesis, we can set

$$a_{ij}v_iv_j = \phi$$

in the given  $x$ -coordinate system and similarly

$$a'_{k\ell}v'_kv'_\ell = \phi' \tag{18.44}$$

in the  $x'$ -coordinate system. In (18.44), we have denoted the as yet unknown transforms of  $a_{ij}$  by  $a'_{ij}$ . Using the transformation law of  $v_i$  as well as the fact that  $\phi = \phi'$  gives us

$$(a_{ij} - R_{ki}R_{\ell j}a'_{k\ell})v_iv_j = 0. \tag{18.45}$$

As a summation is implied over  $i$  and  $j$ , we cannot infer directly that the coefficients of  $v_iv_j$  vanish. Instead, we successively choose components  $(v_1, v_2, v_3, \dots)$  as  $(1, 0, 0, \dots)$  and  $(0, 1, 0, \dots)$ , etc., to get

$$a_{11} - R_{k1}R_{\ell 1}a'_{k\ell} = 0, \quad a_{22} - R_{k2}R_{\ell 2}a'_{k\ell} = 0, \dots \tag{18.46}$$

These results imply that the terms  $a_{ij}$  with  $i = j$  obey the transformation law of second-order tensors. Nevertheless, it tells us nothing about the terms

involving  $a_{ij}$  with  $i \neq j$ . To further examine this point, we set components as  $v_1 \neq 0$ ,  $v_2 \neq 0$ , and  $v_i = 0$  for other  $i$ . Then, (18.45) becomes

$$(a_{11} - R_{k1}R_{\ell1}a'_{k\ell})v_1v_1 + (a_{12} - R_{k1}R_{\ell2}a'_{k\ell})v_1v_2 \\ + (a_{21} - R_{k2}R_{\ell1}a'_{k\ell})v_2v_1 + (a_{22} - R_{k2}R_{\ell2}a'_{k\ell})v_2v_2 = 0.$$

Owing to (18.46), we find that the coefficients of  $v_1v_1$  and  $v_2v_2$  vanish. Furthermore, since

$$R_{k1}R_{\ell2}a'_{k\ell} = R_{k2}R_{\ell1}a'_{\ell k}$$

is simply a relabeling of the indices  $k$  and  $\ell$ , we see that

$$[(a_{12} + a_{21}) - (a'_{k\ell} + a'_{\ell k})R_{k2}R_{\ell1}]v_1v_2 = 0.$$

Thus, choosing  $v_1 = 1$  and  $v_2 = 1$  gives us

$$a_{12} + a_{21} = (a'_{k\ell} + a'_{\ell k})R_{k2}R_{\ell1}.$$

Again, this process may be repeated to yield

$$a_{ij} + a_{ji} = (a'_{k\ell} + a'_{\ell k})R_{kj}R_{\ell i},$$

i.e.,

$$a'_{k\ell} + a'_{\ell k} = R_{kj}R_{\ell i}(a_{ij} + a_{ji}).$$

This is indeed the transformation law of a second-order tensor, but it refers to  $a_{ij} + a_{ji}$ , i.e., the symmetric part of  $2a_{ij}$ , and not to  $a_{ij}$  as such. Accordingly, the quotient theorem for this case must be stated as follows.

♠ **Quotient theorem for two-subscripted quantities:**

Suppose a set of  $n^2$  quantities  $a_{ij}$  to be such that for a vector  $\mathbf{v}$  and for any rotated system, the sum  $a_{ij}v_iv_j$  is a scalar. Then the symmetric parts  $(a_{ij} + a_{ji})/2$  of  $a_{ij}$  are the components of a second-order tensor.

*Remark.*

1. If in addition to the above hypothesis, we are given that the  $a_{ij}$  are symmetric, then the  $a_{ij}$  themselves are the components of a second-order tensor.
2. Nothing can be inferred about the tensorial character of the anti-symmetric part of  $a_{ij}$  from the above hypothesis that because part contributes nothing to the scalar  $\phi$ , as seen from

$$(a_{ij} - a_{ji})v_iv_j = a_{ij}v_iv_j - a_{ji}v_iv_j = a_{ij}v_iv_j - a_{ij}v_jv_i = 0,$$

where in the last step the indices  $i$  and  $j$  are interchanged.

*Example* Using the quotient theorem, we show that the two-subscripted quantities  $a_{ij}$  given by

$$[a_{ij}] = \begin{bmatrix} (x_2)^2 & -x_1x_2 \\ -x_1x_2 & (x_1)^2 \end{bmatrix}$$

are the components of a second-order tensor. Note first that  $a_{ij} = a_{ji}$  and that the outer product  $x_kx_\ell$  is a second-order tensor. Contracting the quantities  $a_{ij}$  with the outer product  $x_kx_\ell$ , we obtain

$$a_{ij}x_ix_j = (x_2)^2(x_1)^2 - x_1x_2x_1x_2 - x_1x_2x_2x_1 + (x_1)^2(x_2)^2 = 0, \quad (18.47)$$

in which the last term, 0, is a zeroth-order tensor. Since (18.47) holds for any rotated coordinate system, we conclude that  $a_{ij}$  is a second-order tensor.

## Exercises

1. Derive the equation  $\nabla \times (\nabla \times \mathbf{v}) = \nabla(\nabla \cdot \mathbf{v}) - \nabla^2 \mathbf{v}$ .

**Solution:** Straightforward calculations yield

$$\begin{aligned} [\nabla \times (\nabla \times \mathbf{v})]_i &= \varepsilon_{ijk} \varepsilon_{klm} \frac{\partial^2 v_m}{\partial x_j \partial x_l} = (\delta_{il} \delta_{jm} - \delta_{im} \delta_{jl}) \frac{\partial^2 v_m}{\partial x_j \partial x_l} \\ &= \frac{\partial^2 v_j}{\partial x_j \partial x_i} - \frac{\partial^2 v_i}{\partial x_j \partial x_j} = \frac{\partial}{\partial x_i} \left( \frac{\partial v_j}{\partial x_j} \right) - \frac{\partial^2 v_i}{\partial x_j \partial x_j} \\ &= \frac{\partial}{\partial x_i} (\nabla \cdot \mathbf{v}) - \nabla^2 v_i = [\nabla(\nabla \cdot \mathbf{v})]_i - [\nabla^2 \mathbf{v}]_i \\ &= [\nabla(\nabla \cdot \mathbf{v}) - \nabla^2 \mathbf{v}]_i. \quad \clubsuit \end{aligned}$$

2. Derive the expression (18.38) using the result (18.37).

**Solution:** We consider the vector products (in the sense of elementary vector calculus) of the transformed basis arrows  $\mathbf{e}'_i$  given by  $\mathbf{e}'_i \times \mathbf{e}'_j = (R_{i\ell} \mathbf{e}_\ell) \times (R_{jm} \mathbf{e}_m) = R_{i\ell} R_{jm} \mathbf{e}_\ell \times \mathbf{e}_m$ . Forming the scalar product with  $\mathbf{e}_n$  yields

$$(\mathbf{e}'_i \times \mathbf{e}'_j) \cdot \mathbf{e}_n = R_{i\ell} R_{jm} (\mathbf{e}_\ell \times \mathbf{e}_m) \cdot \mathbf{e}_n,$$

where on the right-hand side only two terms survive for each fixed value of  $n$  since

$$(\mathbf{e}_\ell \times \mathbf{e}_m) \cdot \mathbf{e}_n = \begin{cases} +1 & \text{if } (\ell, m, n) = (1, 2, 3), (2, 3, 1), (3, 1, 2), \\ -1 & \text{if } (\ell, m, n) = (2, 1, 3), (3, 2, 1), (1, 3, 2), \\ 0 & \text{otherwise.} \end{cases}$$

(Here we assume that the coordinate systems associated with  $\{\mathbf{e}_i\}$  and  $\{\mathbf{e}'_j\}$  are both right-handed.) Hence, we have

$$(\mathbf{e}'_i \times \mathbf{e}'_j) \cdot \mathbf{e}_n = R_{i\ell}R_{jm} - R_{im}R_{j\ell}, \quad (18.48)$$

where  $\ell, m, n$  is a cyclic permutation of 1, 2, 3. Moreover, since

$$\mathbf{e}'_i \times \mathbf{e}'_j = \mathbf{e}'_k = R_{kr}\mathbf{e}_r, \quad (18.49)$$

it follows from (18.48) and (18.49) that

$$R_{kr}\mathbf{e}_r \cdot \mathbf{e}_n = R_{kr}\delta_{rn} = R_{kn} = R_{i\ell}R_{jm} - R_{im}R_{j\ell},$$

where again  $i, j, k$  and  $\ell, m, n$  are both cyclic permutations of 1, 2, 3. If  $\{\mathbf{e}'_\ell\}$  is left-handed, a similar procedure yields

$$R_{kn} = -(R_{i\ell}R_{jm} - R_{im}R_{j\ell}).$$

Substituting these results into (18.37), we finally arrive at the conclusion that

$$w'_k = |R|R_{kn}w_n,$$

which is a transformation law for a pseudovector. ♣

3. Show that the process of contraction of an  $N$ th-order tensor produces another tensor of order  $N - 2$ .

**Solution:** Let  $T_{ij\dots l\dots m\dots k}$  be the components of an  $N$ th-order tensor; then

$$T'_{ij\dots l\dots m\dots k} = R_{ip}R_{jq} \cdots R_{lr} \cdots R_{ms} \cdots R_{kn}T_{pq\dots r\dots s\dots n}.$$

Thus if, e.g., we make the two subscripts  $l$  and  $m$  equal and sum over all the values of these subscripts, we obtain

$$\begin{aligned} T'_{ij\dots l\dots m\dots k} &= R_{ip}R_{jq} \cdots R_{lr} \cdots R_{ms} \cdots R_{kn}T_{pq\dots r\dots s\dots n} \\ &= R_{ip}R_{jq} \cdots \delta_{rs} \cdots R_{kn}T_{pq\dots r\dots s\dots n} \\ &= R_{ip}R_{jq} \cdots R_{kn}T_{pq\dots r\dots r\dots n}, \end{aligned}$$

showing that  $T_{ij\dots l\dots l\dots k}$  are the components of a (different) Cartesian tensor of order  $N - 2$ . ♣

## 18.5 Applications in Physics and Engineering

This section is devoted to illustrations of physical applications of second- and higher-order Cartesian tensors. We start with an example from mechanics and follow that by examples from electromagnetism and elasticity.

### 18.5.1 Inertia Tensor

Consider a collection of rigidly connected particles, wherein the  $\alpha$ th particle has mass  $m^{(\alpha)}$  and is positioned at  $\mathbf{r}^{(\alpha)}$  with respect to the origin  $O$ . Suppose that the rigid assembly is rotating about an axis through  $O$  with **angular velocity**  $\boldsymbol{\omega}$ . The **angular momentum**  $\mathbf{J}$  of the assembly is given by

$$\mathbf{J} = \sum_{\alpha} \left( \mathbf{r}^{(\alpha)} \times \mathbf{p}^{(\alpha)} \right).$$

Here  $\mathbf{p}^{(\alpha)} = m^{(\alpha)} \dot{\mathbf{r}}^{(\alpha)}$  and  $\dot{\mathbf{r}}^{(\alpha)} = \boldsymbol{\omega} \times \mathbf{r}^{(\alpha)}$  for any  $\alpha$  whose components are expressed in subscript form as

$$p_k^{(\alpha)} = m^{(\alpha)} \dot{x}_k^{(\alpha)} \quad \text{and} \quad \dot{x}_k^{(\alpha)} = \varepsilon_{klm} \omega_l x_m^{(\alpha)}.$$

Thus we obtain

$$\begin{aligned} J_i &= \sum_{\alpha} \sum_{j,k} \varepsilon_{ijk} x_j^{(\alpha)} \dot{p}_k^{(\alpha)} = \sum_{\alpha} \sum_{j,k,l,m} m^{(\alpha)} \varepsilon_{ijk} x_j^{(\alpha)} \varepsilon_{klm} \omega_l x_m^{(\alpha)} \\ &= \sum_{\alpha} \sum_{j,l} m^{(\alpha)} (\delta_{il} \delta_{jm} - \delta_{im} \delta_{jl}) x_j^{(\alpha)} x_m^{(\alpha)} \omega_l \\ &= \sum_{\alpha} \sum_l m^{(\alpha)} \left[ \left( r^{(\alpha)} \right)^2 \delta_{il} - x_i^{(\alpha)} x_l^{(\alpha)} \right] \omega_l \equiv \sum_l I_{il} \omega_l, \end{aligned} \quad (18.50)$$

with the definition

$$I_{il} = \sum_{\alpha} m^{(\alpha)} \left[ \left( r^{(\alpha)} \right)^2 \delta_{il} - x_i^{(\alpha)} x_l^{(\alpha)} \right]. \quad (18.51)$$

The set of quantities  $I_{il}$  forms a symmetric second-order Cartesian tensor; the symmetric property expressed by  $I_{il} = I_{li}$  follows readily from (18.51). The fact that the  $I_{il}$  form tensors can be proved by applying the quotient rule (see Sect. 18.4.6) to equation (18.50), wherein  $J_i$  and  $\omega_l$  are vectors. The tensor  $I_{il}$  is called the **inertia tensor** of the assembly with respect to  $O$ . As evident from (18.51),  $I_{il}$  depends only on the distribution of mass in the assembly and not on the direction or magnitude of the angular velocity of the assembly,  $\boldsymbol{\omega}$ .

If a continuous rigid body is considered,  $m^{(\alpha)}$  is replaced by the mass distribution  $\rho(\mathbf{r})$  and the summation  $\sum_{\alpha}$  by the integral of  $\int dV$  over the volume of the whole body. When expanded in Cartesian coordinates, the inertia tensor of a continuous body would have the form

$$\mathbf{I} = [I_{ij}] = \begin{pmatrix} \int (y^2 + z^2) \rho dV & - \int xy \rho dV & - \int zx \rho dV \\ - \int xy \rho dV & \int (z^2 + x^2) \rho dV & - \int yz \rho dV \\ - \int zx \rho dV & - \int yz \rho dV & \int (x^2 + y^2) \rho dV \end{pmatrix}.$$

The diagonal elements of this tensor are called the **moments of inertia** and the off-diagonal elements without the negative signs are known as the **products of inertia**.

It is possible to show that the **kinetic energy**  $K$  of the rotating system is given by  $K = \frac{1}{2} I_{jl} \omega_j \omega_l$ , which is a scalar obtained by twice contracting

the vector  $\omega_j$  with the inertia tensor  $I_{jl}$ . In fact, an argument parallel to that leading to (18.50) yields

$$\begin{aligned} K &= \frac{1}{2} \sum_{\alpha} m^{(\alpha)} \dot{\mathbf{r}}^{(\alpha)} \cdot \dot{\mathbf{r}}^{(\alpha)} = \frac{1}{2} \sum_{\alpha} m^{(\alpha)} \sum_{j,k,l,m} \varepsilon_{ijk} \omega_j x_k^{(\alpha)} \varepsilon_{ilm} \omega_l x_m^{(\alpha)} \\ &= \frac{1}{2} \sum_{j,l} I_{jl} \omega_j \omega_l. \end{aligned}$$

This shows that the kinetic energy of the rotating body can be expressed as a scalar obtained by twice contracting the vector  $\omega_j$  with the inertia tensor  $I_{jl}$ . Alternatively, since  $J_j = I_{jl} \omega_l$ , the kinetic energy may be written as  $K = \frac{1}{2} J_j \omega_j$ .

### 18.5.2 Tensors in Electromagnetism in Solids

**Magnetic susceptibility** and **electric conductivity** are also examples of physical quantities represented by second-order tensors. For the former, we have the standard expression

$$M_i = \sum_j \chi_{ij} H_j, \quad (18.52)$$

where  $\mathbf{M}$  is the magnetic moment per unit volume and  $\mathbf{H}$  is the magnetic field. Similarly, for the case of electric conductivity, we can write

$$j_i = \sum_j \sigma_{ij} E_j. \quad (18.53)$$

Here, the current density  $\mathbf{j}$  (current per unit perpendicular area) is related to the electric field  $\mathbf{E}$ . In both cases, we have a vector on the left-hand side and the contraction of a second-order tensor with another vector on the right-hand side.

For isotropic media, the vector  $\mathbf{M}$  is parallel to  $\mathbf{H}$  and, similarly, the vector  $\mathbf{j}$  is parallel to  $\mathbf{E}$ . Thus, the above tensors satisfy  $\chi_{ij} = \chi \delta_{ij}$  and  $\sigma_{ij} = \sigma \delta_{ij}$ , respectively, resulting in  $\mathbf{M} = \chi \mathbf{H}$  and  $\mathbf{j} = \sigma \mathbf{E}$ . However, for anisotropic materials such as crystals, the magnetic susceptibility and electric conductivity may be different along different crystal axes, thus making  $\chi_{ij}$  and  $\sigma_{ij}$  general second-order tensors (usually symmetric).

### 18.5.3 Electromagnetic Field Tensor

All the tensors that we have considered in this chapter so far relate to the three dimensions of space and they are defined as having a certain transformation property under spatial rotations. In this subsection, we shall have the occasion

to use a tensor in the four dimensions of relativistic space-time; the tensor is the electromagnetic field tensor  $F_{\mu\nu}$ .

Recall that an **electromagnetic field** in free space is governed by the **Maxwell equations**, which take the form

$$\begin{aligned}\nabla \cdot \mathbf{B} &= 0, & \nabla \cdot \mathbf{E} &= 4\pi k_1 \rho, \\ \nabla \times \mathbf{B} &= 4\pi k_2 \mathbf{J} + \frac{k_2}{k_1} \frac{\partial \mathbf{E}}{\partial t}, & \nabla \times \mathbf{E} &= -k_3 \frac{\partial \mathbf{B}}{\partial t}.\end{aligned}$$

Here  $\mathbf{E}$  is the electric field intensity,  $\mathbf{B}$  is the magnetic induction,  $\rho$  is the charge density, and  $\mathbf{J}$  is the current density. There are several ways of defining the values of constants  $k_i$  ( $i = 1, 2, 3$ ); indeed, their values depends on which system of unit we use. Typical examples are listed in Table 18.1.

The Maxwell equations take on a particularly simple and elegant form on introducing the **electromagnetic field tensor**  $F_{\mu\nu}$  defined as

$$F_{\mu\nu} = \partial_\nu A_\mu - \partial_\mu A_\nu. \quad (18.54)$$

Here,  $A_\mu = (\phi/c, -\mathbf{A})$  is called a **four potential**, determined by the scalar potential  $\phi$  and the vector potential  $\mathbf{A}$  that generate the fields  $\mathbf{B} = \nabla \times \mathbf{A}$  and  $\mathbf{E} = -\nabla\phi - \partial\mathbf{A}/\partial t$ . The symbol  $\partial_\mu$  in (18.54) denotes the partial derivatives with respect to the  $\mu$ th coordinate. Straightforward calculations yield

$$[F_{\mu\nu}] = \begin{bmatrix} 0 & E^1/c & E^2/c & E^3/c \\ -E^1/c & 0 & -B^3 & B^2 \\ -E^2/c & B^3 & 0 & -B^1 \\ -E^3/c & -B^2 & B^1 & 0 \end{bmatrix}, \quad (18.55)$$

where  $\mathbf{E} = (E^1, E^2, E^3)$  and  $\mathbf{B} = (B^1, B^2, B^3)$ . We also introduce another relevant tensor defined by

$$[F^{\mu\nu}] = \begin{bmatrix} 0 & -E^1/c & -E^2/c & -E^3/c \\ E^1/c & 0 & -B^3 & B^2 \\ E^2/c & B^3 & 0 & -B^1 \\ E^3/c & -B^2 & B^1 & 0 \end{bmatrix}, \quad (18.56)$$

in which  $\mu$  and  $\nu$  are *superscripts* in opposed to (18.55), where they are *subscripts*. As a result, we can see that the Maxwell equations are equivalent

**Table 18.1.** Values of the constants  $k_i$  ( $i = 1, 2, 3$ ) in the Maxwell equations.  $\mu_0$ ,  $\varepsilon_0$  and  $c$  are the permeability, permittivity, and speed of light in vacuum, respectively

System of Unit	$k_1$	$k_2$	$k_3$
MKSA	$1/(4\pi\varepsilon_0)$	$\mu_0/(4\pi)$	1
CGS-esu	1	$1/c^2$	1
CGS-emu	$c^2$	1	1
CGS-Gauss	1	$1/c$	$1/c$

to the following two field equations:

$$\sum_{\nu} \partial_{\nu} F^{\mu\nu} = \mu_0 j^{\mu},$$

$$\partial_{\sigma} F_{\mu\nu} + \partial_{\mu} F_{\nu\sigma} + \partial_{\nu} F_{\sigma\mu} = 0,$$

where  $j^{\mu} = (\rho c, \mathbf{J})$  is the **four-current density**.

**Remark.** The distinction between superscripts and subscripts on the symbol  $F$  shown in (18.55) and (18.56), respectively, is clarified in Chap. 19, which deals with non-Cartesian tensor calculus.

#### 18.5.4 Elastic Tensor

Thus so far, we have focused on the physical applications of second-order tensors, which relate two vectors. Now, we extend this idea to a situation where a fourth-order tensor relates two physical second-order tensors. Such relationships commonly occur in **elasticity theory**. In the framework of this theory, the local deformation of an elastic body at any interior point  $P$  is described by a second-order symmetric tensor  $e_{ij}$  called the **strain tensor**, which is given by

$$e_{ij} = \frac{1}{2} \left( \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right),$$

where  $\mathbf{u}$  is the **displacement vector** describing the strain of a small volume element. Similarly, we can describe the stress in the body at  $P$  by a second-order symmetric **stress tensor**  $p_{ij}$ ; the quantity  $p_{ij}$  is the  $x_j$ -component of the stress vector acting across a plane through  $P$ , whose normal lies in the  $x_i$ -direction. A generalization of **Hooke's law** that relates the stress and strain tensors is

$$p_{ij} = \sum_{k,l} c_{ijkl} e_{kl}, \quad (18.57)$$

where  $c_{ijkl}$  is a fourth-order Cartesian tensor.

Specifically, for an isotropic medium, we must have an isotropic tensor for  $c_{ijkl}$ ; the most general fourth-order isotropic tensor is

$$c_{ijkl} = \lambda \delta_{ij} \delta_{kl} + \eta \delta_{ik} \delta_{jl} + \nu \delta_{il} \delta_{jk}.$$

Substituting this into (18.57) yields

$$p_{ij} = \lambda \delta_{ij} \sum_k e_{kk} + \eta e_{ij} + \nu e_{ji}. \quad (18.58)$$

Note that  $e_{ij}$  is symmetric. Hence, if we write  $\eta + \nu = 2\mu$ , (18.58) takes the conventional form

$$p_{ij} = \lambda \sum_k e_{kk} \delta_{ij} + 2\mu e_{ij},$$

in which  $\lambda$  and  $\mu$  are known as **Lamé constants**.

## Non-Cartesian Tensors

---

**Abstract** Having discussed tensor theory based on Cartesian coordinates, we now move on to its counterpart, i.e., tensors described by curvilinear coordinate systems. The use of a curvilinear coordinate system endows the tensor calculus with the properties of “covariance” (Sect. 19.1.3) and “contravariance” (Sect. 19.1.4), both of which are new concepts originating from the nonorthogonality of the coordinate axes.

### 19.1 Curvilinear Coordinate Systems

#### 19.1.1 Local Basis Vectors

We have thus far restricted our attention to the study of Cartesian tensors, where, from a practical stand point, only rigid rotations of axes (proper and/or improper) are taken into account as coordinate transformations. However, we must free ourselves from this restriction and develop the tensor calculus in terms of **curvilinear coordinate systems**. In advanced mathematical physics, we often have to deal with tensor analysis on curved surfaces (or more abstract manifolds) on which orthonormal coordinate systems cannot be defined, and in such cases the theory developed thus far is entirely inadequate. This means that we have to formulate tensors and their transformations in terms of general curvilinear coordinate systems.

To begin with, we review some properties of general curvilinear coordinates. Suppose that the position of an arbitrary point  $P$  in a three-dimensional space has Cartesian coordinates  $x, y, z$ . In general, this position may be expressed in terms of three curvilinear coordinates  $u_1, u_2, u_3$ , which are functions of  $x, y, z$  as explicitly represented by

$$\begin{aligned} u_1 &= u_1(x, y, z), \\ u_2 &= u_2(x, y, z), \\ u_3 &= u_3(x, y, z). \end{aligned}$$

We denote by  $\mathbf{r}$  the position arrow connecting the origin  $O$  and the point  $P$ . Obviously, the direction and magnitude of the arrow depend on the coordinates of  $P$ , which are symbolized by

$$\mathbf{r} = \mathbf{r}(u_1, u_2, u_3).$$

We now consider the partial derivative of  $\mathbf{r}$  with respect to  $u_i$ , i.e.,

$$\mathbf{e}_i \equiv \frac{\partial \mathbf{r}}{\partial u_i}. \quad (19.1)$$

From the definition, the vectors  $\mathbf{e}_i$  are directed along the corresponding coordinate lines at the point  $P$ . As a result, an infinitesimal vector displacement  $d\mathbf{r}$  in curvilinear coordinates is given by

$$d\mathbf{r} = \frac{\partial \mathbf{r}}{\partial u_i} du_i = \mathbf{e}_i du_i,$$

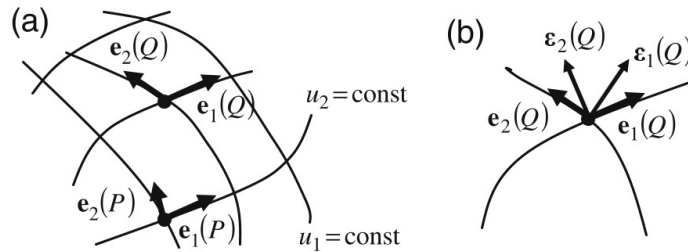
where the summation convention is employed. The vectors  $\mathbf{e}_i$  are referred to as **local basis vectors**. (In precise terminology, they are called **covariant local basis vectors**, as explained later.)

It is obvious from (19.1) that the vectors  $\mathbf{e}_i$  are functions of the curvilinear coordinates  $u_i$ , namely,  $\mathbf{e}_i = \mathbf{e}_i(u_1, u_2, u_3)$ . This implies that the directions and magnitudes of the  $\mathbf{e}_i$  vary from point to point in the space considered, which is in contrast to the case of a Cartesian coordinate system, where the basis vectors are spatially independent. Spatial dependence of basis vectors is actually one of the most important properties of curvilinear coordinate systems.

Another notable property of curvilinear coordinate systems is the fact that they allow us to define another useful set of three vectors at  $P$  as

$$\boldsymbol{\varepsilon}_i \equiv \nabla u_i.$$

Clearly the direction of  $\boldsymbol{\varepsilon}_i$  is normal to the surface  $u_i = \text{const}$ ; thus being different from the directions of any vectors  $\mathbf{e}_i$  ( $i = 1, 2, 3$ ) in general (see Fig. 19.1). Therefore, at each point  $P$  in a curvilinear coordinate system, there exist two sets of basis vectors defined by



**Fig. 19.1.** (a) Spatial dependence of  $\mathbf{e}_i$  in the curvilinear coordinate system. (b) Difference between  $\mathbf{e}_i$  and  $\boldsymbol{\varepsilon}_i$

$$\mathbf{e}_i = \frac{\partial \mathbf{r}}{\partial u_i} \quad \text{and} \quad \boldsymbol{\varepsilon}_i = \nabla u_i. \quad (19.2)$$

In the tensor analysis, literature of the set of vectors  $\boldsymbol{\varepsilon}_i$  introduced above is denoted by  $\mathbf{e}^i$ , the index being placed as a superscript to distinguish it from the first set of vectors  $\mathbf{e}_i$ . Relating to the notation above, we introduce a **modified summation convention** as follows: if we find a lower-case alphabetic index that appears twice, once as a subscript and once as a superscript, we sum over all the values that the index can take. In this convention, the curvilinear coordinates are denoted by  $u^1, u^2, u^3$ , with the index raised (see the remark in Sect. 19.1.3), to arrive at the following definition.

♠ **Local basis vectors:**

A curvilinear coordinate system is characterized by two sets of three vectors  $\{\mathbf{e}_i\}$  and  $\{\mathbf{e}^i\}$  defined by

$$\mathbf{e}_i = \frac{\partial \mathbf{r}}{\partial u^i} \quad \text{and} \quad \mathbf{e}^i = \nabla u^i.$$

Here, the  $\mathbf{e}_i$  are referred to as the **covariant local basis vectors**, and the  $\mathbf{e}^i$  as the **contravariant local basis vectors**.

The prefix “local” emphasizes the fact that the lengths and orientations of these basis vectors vary from point to point in the space; this fact is explicitly represented by

$$\mathbf{e}_i = \mathbf{e}_i(u^1, u^2, u^3) \quad \text{and} \quad \mathbf{e}^j = \mathbf{e}^j(u^1, u^2, u^3).$$

For the sake of conciseness, we omit the prefix in the subsequent discussions and use the terms **contravariant** (or **covariant**) **basis vectors**, bearing the locality in mind.

*Remark.*

1. In common practice indices that represent *contravariant* character are placed as superscripts and those indicating *covariant* character as subscripts.
2. For Cartesian coordinate systems, the two sets of basis vectors  $\mathbf{e}_i$  and  $\mathbf{e}^i$  are identical and, hence, there is no need to differentiate between contravariance and covariance.
3. In derivatives such as  $\partial \mathbf{r} / \partial u^i$ , the  $i$  is considered as a *subscript*.

### 19.1.2 Reciprocity Relations

Generally the covariant basis vectors  $\mathbf{e}_1$ ,  $\mathbf{e}_2$ , and  $\mathbf{e}_3$  are neither of unit length nor are they orthogonal to each other; this is also true for the contravariant

basis vectors,  $\mathbf{e}^1$ ,  $\mathbf{e}^2$ , and  $\mathbf{e}^3$ . Nevertheless, the sets  $\{\mathbf{e}_i\}$  and  $\{\mathbf{e}^j\}$  still have an important property as stated below.

♠ **Reciprocity relations:**

The sets of contravariant and covariant local basis vectors  $\{\mathbf{e}_i\}$  and  $\{\mathbf{e}^j\}$  satisfy the reciprocity relations such that

$$\mathbf{e}_i \cdot \mathbf{e}^j = \delta_i^j, \quad (19.3)$$

where the scalar product of the vectors is taken in the sense of elementary vector calculus.

*Proof* By using Cartesian representation, we have

$$\begin{aligned} \mathbf{e}_i \cdot \mathbf{e}^j &= \frac{\partial \mathbf{r}}{\partial u^i} \cdot \nabla u^j = \left( \frac{\partial x}{\partial u^i}, \frac{\partial y}{\partial u^i}, \frac{\partial z}{\partial u^i} \right) \cdot \left( \frac{\partial u^j}{\partial x}, \frac{\partial u^j}{\partial y}, \frac{\partial u^j}{\partial z} \right) \\ &= \frac{\partial x}{\partial u^i} \frac{\partial u^j}{\partial x} + \frac{\partial y}{\partial u^i} \frac{\partial u^j}{\partial y} + \frac{\partial z}{\partial u^i} \frac{\partial u^j}{\partial z} \\ &= \frac{\partial u^j}{\partial u^i} = \delta_i^j. \quad \clubsuit \end{aligned}$$

*Remark.* The reciprocity relation (19.3) implies that each covariant (or contravariant) basis vector  $\mathbf{e}_i$  (or  $\mathbf{e}^i$ ) is perpendicular to all contravariant (or covariant) basis vectors  $\mathbf{e}_k$  (or  $\mathbf{e}^k$ ) except  $k = i$ . For instance,  $\mathbf{e}_1$  is perpendicular to  $\mathbf{e}^2$  and  $\mathbf{e}^3$ , but not to  $\mathbf{e}^1$  in general. To be precise, the vectors  $\mathbf{e}_1$  and  $\mathbf{e}^1$  make an angle  $\theta$  that satisfies

$$\mathbf{e}_1 \cdot \mathbf{e}^1 = |\mathbf{e}_1| |\mathbf{e}^1| \cos \theta = 1,$$

where  $|\mathbf{e}_1| \neq 1$  and  $|\mathbf{e}^1| \neq 1$ .

### 19.1.3 Transformation Law of Covariant Basis Vectors

We are now in a position to discuss the concept of general transformations from one coordinate system,  $u^1, u^2, u^3$ , to another,  $u'^1, u'^2, u'^3$ . A coordinate transformation is described by using the three equations

$$u'^i = u'^i(u^1, u^2, u^3), \quad (19.4)$$

for  $i = 1, 2, 3$ , in which the new coordinates  $u'^i$  can be arbitrary functions of the old ones  $u^i$ . We assume that the transformation can be inverted, so that we can write the old coordinates in terms of the new ones as

$$u^i = u^i(u'^1, u'^2, u'^3).$$

We now formulate the transformation law of basis vectors. The two sets of basis vectors in the new coordinate system are given by

$$\mathbf{e}'_i = \frac{\partial \mathbf{r}}{\partial u'^i} \quad \text{and} \quad \mathbf{e}'^i = \nabla u'^i. \quad (19.5)$$

Using the chain rule, we find that the first set of basis vectors yields

$$\mathbf{e}'_i = \frac{\partial \mathbf{r}}{\partial u^j} \frac{\partial u^j}{\partial u'^i} = \frac{\partial u^j}{\partial u'^i} \mathbf{e}_j. \quad (19.6)$$

This describes the transformation behavior of the local covariant basis vectors from the unprimed one  $\mathbf{e}_j$  to the primed one  $\mathbf{e}'_i$  under the coordinate transformation (19.4). Note that the partial derivatives as well as the basis vectors in (19.6) vary from point to point. Hence, relation (19.6) is valid under the condition that all terms involved are evaluated at the same point  $P$  in the space being considered.

In the same manner, it follows that

$$\mathbf{e}_k = \frac{\partial \mathbf{r}}{\partial u^j} = \frac{\partial \mathbf{r}}{\partial u'^\ell} \frac{\partial u'^\ell}{\partial u^k} = \frac{\partial u'^\ell}{\partial u^k} \mathbf{e}'_\ell.$$

We thus have proved the following theorem:

♠ **Transformation law of covariant basis vectors:**

The sets of local covariant basis vectors  $\{\mathbf{e}_i(u^j)\}$  and  $\{\mathbf{e}'_k(u'^\ell)\}$  associated with two different curvilinear coordinate systems are related at a point  $P$  by

$$\mathbf{e}'_i = \frac{\partial u^j}{\partial u'^i} \mathbf{e}_j \quad \text{and} \quad \mathbf{e}_k = \frac{\partial u'^\ell}{\partial u^k} \mathbf{e}'_\ell, \quad (19.7)$$

in which the partial derivatives are to be evaluated at  $P$ .

**Remark.** Observe that in all the mathematical expressions above (and below), the summation convention is applied to the indices that are repeated in one term as both a subscript and a superscript. Indeed, it was to satisfy this summation convention that the coordinates were written as  $u^i$  rather than  $u_i$ .

### 19.1.4 Transformation Law of Contravariant Basis Vectors

Next we consider the transformation law of the contravariant basis vectors  $\mathbf{e}^i = \nabla u^i$ . Recall that in terms of a rectangular Cartesian coordinate system, the operator  $\nabla$  is expressed as

$$\nabla = \mathbf{i} \frac{\partial}{\partial x} + \mathbf{j} \frac{\partial}{\partial y} + \mathbf{k} \frac{\partial}{\partial z},$$

where  $\mathbf{i}, \mathbf{j}, \mathbf{k}$  are mutually orthogonal basis vectors of unit length. It then, follows that

$$\mathbf{e}'^k = \nabla u'^k = \mathbf{i} \frac{\partial u'^k}{\partial x} + \mathbf{j} \frac{\partial u'^k}{\partial y} + \mathbf{k} \frac{\partial u'^k}{\partial z},$$

in which the first partial derivative reads

$$\frac{\partial u'^k}{\partial x} = \frac{\partial u^i}{\partial x} \frac{\partial u'^k}{\partial u^i}$$

and other derivatives are written in the same way. Hence, we have

$$\mathbf{e}'^k = \left( \mathbf{i} \frac{\partial u^i}{\partial x} + \mathbf{j} \frac{\partial u^i}{\partial y} + \mathbf{k} \frac{\partial u^i}{\partial z} \right) \frac{\partial u'^k}{\partial u^i} = (\nabla u^i) \frac{\partial u'^k}{\partial u^i} = \frac{\partial u'^k}{\partial u^i} \mathbf{e}^i.$$

Similarly, we have

$$\mathbf{e}^j = \nabla u^j = \left( \mathbf{i} \frac{\partial u'^\ell}{\partial x} + \mathbf{j} \frac{\partial u'^\ell}{\partial y} + \mathbf{k} \frac{\partial u'^\ell}{\partial z} \right) \frac{\partial u^j}{\partial u'^\ell} = \frac{\partial u^j}{\partial u'^\ell} \mathbf{e}'^\ell.$$

These results are summarized as follows:

♠ **Transformation law of contravariant basis vectors:**

The two sets of local contravariant basis vectors  $\{\mathbf{e}^i(u^j)\}$  and  $\{\mathbf{e}'^k(u'^\ell)\}$  are related at a point  $P$  by

$$\mathbf{e}'^k = \frac{\partial u'^k}{\partial u^i} \mathbf{e}^i \quad \text{and} \quad \mathbf{e}^\ell = \frac{\partial u^\ell}{\partial u'^j} \mathbf{e}'^j, \quad (19.8)$$

where the partial derivatives are again to be evaluated at  $P$ .

It should be emphasized again that, owing to the summation convention, the repeated indices in (19.8) appear once as a superscript and once as a subscript

### 19.1.5 Components of a Vector

Given the two bases  $\mathbf{e}_i$  and  $\mathbf{e}^i$ , we may express a general geometric arrow  $\mathbf{a}$  (i.e., a vector  $\mathbf{a}$ ) equally well in terms of either basis as follows:

$$\mathbf{a} = a^1 \mathbf{e}_1 + a^2 \mathbf{e}_2 + a^3 \mathbf{e}_3 = a^i \mathbf{e}_i,$$

$$\mathbf{a} = a_1 \mathbf{e}^1 + a_2 \mathbf{e}^2 + a_3 \mathbf{e}^3 = a_i \mathbf{e}^i.$$

The  $a^i$  are called the **contravariant components** of the vector  $\mathbf{a}$  and the  $a_i$  the **covariant components**. Both kinds of components  $a^i$  and  $a_j$  describe the same vector  $\mathbf{a}$ , but they are associated with different basis vectors  $\mathbf{e}_i$  and  $\mathbf{e}^j$ , respectively. In plain words, a vector assigned at a point in a curvilinear coordinate system has two different expressions; say,  $(a^1, a^2, a^3)$  and  $(a_1, a_2, a_3)$  for the same vector  $\mathbf{a}$ . The tensorial characters of the two kinds of components are inherently different from each other, as we shall see in subsequent discussions.

For any vector  $\mathbf{a}$ , the two kinds of components  $a^i$  and  $a_i$  are readily obtained by forming the scalar products,

$$\mathbf{a} \cdot \mathbf{e}^i = a^j \mathbf{e}_j \cdot \mathbf{e}^i = a^j \delta_j^i = a^i$$

and

$$\mathbf{a} \cdot \mathbf{e}_i = a^j \mathbf{e}^j \cdot \mathbf{e}_i = a_j \delta_i^j = a_i,$$

where we have used the reciprocity relation (19.3). Furthermore, using the transformation law of  $\mathbf{e}_i$  given in (19.7) gives us

$$\mathbf{a} = a'^i \mathbf{e}'_i = a^j \mathbf{e}_j = a^j \frac{\partial u'^i}{\partial u^j} \mathbf{e}'_i. \quad (19.9)$$

This provides the transformation law of the contravariant components of a vector such that

$$a'^i = \frac{\partial u'^i}{\partial u^j} a^j. \quad (19.10)$$

This relation is, in fact, the defining property for a set of quantities  $a^i$  to form the contravariant components of a vector. The formal statement is given below.

♠ **Contravariant component of a vector:**

Quantities  $a_i$  associated with a point  $P$  are said to be the contravariant components of a vector if these, quantities transform through the equation

$$a'^i = \frac{\partial u'^i}{\partial u^j} a^j, \quad (19.11)$$

where the partial derivatives are evaluated at  $P$ .

**Remark.** It might occur that a given ordered set of quantities  $a^k$  associated with a point  $P$  has nothing to do with a vector; only those sets satisfying the transformation law (19.11) serve as (contravariant) components of a vector.

Analogously to the case of (19.9), it follows from the identity for an arbitrary vector  $\mathbf{a}$ ,

$$\mathbf{a} = a'_i \mathbf{e}'^i = a_j \mathbf{e}^j = a_j \frac{\partial u^j}{\partial u'^i} \mathbf{e}'^i,$$

that the transformation law of covariant components yields

$$a'_i = \frac{\partial u^j}{\partial u'^i} a_j. \quad (19.12)$$

Again we take this result as the defining property of the covariant components of a vector.

♠ **Covariant components of a vector:**

Quantities  $a_i$  associated with a point  $P$  are said to be the covariant components of a vector if those quantities transform through the equation

$$a'_k = \frac{\partial u^\ell}{\partial u'^k} a_\ell, \quad (19.13)$$

where the partial derivatives are evaluated at  $P$ .

*Remark.* Other textbooks may use the expression “**contravariant** (or **covariant**) **vector**,” which is a distinctly different concept from a vector  $\mathbf{a}$  or its components  $a^i$  (or  $a_i$ ) that we have just defined. Say, rather, that a contravariant vector is a *collection* of ordered triples,

$$\left\{ (a^1, a^2, a^3), (a'^1, a'^2, a'^3), (a''^1, a''^2, a''^3), \dots \right\},$$

in which all the ordered triples consist of contravariant components of the *same* vector  $\mathbf{a}$  associated with different coordinate systems. We should make sure that a contravariant (or covariant) vector is not expressed by a geometric arrow as is done for a vector.

### 19.1.6 Components of a Tensor

We now define geometric objects of the contravariant class, which are more complicated in character than vectors and begin with the following:

♠ **Contravariant component of a tensor:**

Index quantities  $T_{jk}$  associated with a point  $P$  are said to be **contravariant components** of a tensor if these quantities transform according to the equation

$$T'^{jk} = T^{\ell m} \frac{\partial u'^j}{\partial u^\ell} \frac{\partial u'^k}{\partial u^m}. \quad (19.14)$$

There is no difficulty in defining covariant tensors of higher orders. For a tensor of second order, e.g., we have the definition below.

♠ **Covariant components of a tensor:**

Index quantities  $T_{jk}$  are said to be covariant components of a second-order tensor if these quantities transform according to the equation

$$T'_{jk} = T_{\ell m} \frac{\partial u^\ell}{\partial u'^j} \frac{\partial u^m}{\partial u'^k}. \quad (19.15)$$

We shall see later that there are many examples of tensors of this kind in physics and engineering. The moment of inertia, the stress of elasticity, and the electromagnetic field are cases in point; if their components in terms of certain coordinate systems are evaluated, they all turn out to obey the transformation law (19.14).

In terminology, all quantities satisfying (19.11), (19.13) and (19.14), (19.15) are called **components of a first-order tensor** and **components of a second-order tensor**, respectively; the order goes as the number of indices attached. The definitions of tensors of higher orders are given through a straightforward generalization of the above. Conversely, we can define a **tensor of zero order**, called a **scalar**, that involves no index so that its single component (i.e., the scalar itself) is constant under any coordinate transformation; namely,

$$T' = T.$$

Such a quantity is called an **invariant**.

**Remark.** For any components of tensors, the number of indices is independent of the number of dimensions of the space considered. The definitions above for vectors, tensors, and scalars are all valid for an arbitrary  $n$ -dimensional space.

### 19.1.7 Mixed Components of a Tensor

Having defined contravariant and covariant components of a tensor, we can now define another class of components, called **mixed components of a tensor**, that involve the two character simultaneously.

♠ **Mixed components of a tensor:**

Index quantities  $T^i_{jk}$  are said to be the mixed components of a tensor of the third order if these quantities transform according to the equation

$$T'^i_{jk} = T^{\ell}_{mn} \frac{\partial u'^i}{\partial u^{\ell}} \frac{\partial u^m}{\partial u'^j} \frac{\partial u^n}{\partial u'^k}.$$

Clearly,  $T^i_{jk}$  transforms contravariantly with respect to the first index  $i$  but covariantly with respect to the other indices  $j$  and  $k$ .

If we consider the components of higher-order tensors in non-Cartesian coordinates, there are even more possibilities. As an example, let us consider a second-order tensor  $\mathbf{T}$ . Using the outer product notation, we may write  $\mathbf{T}$  in three different ways:

$$\mathbf{T} = T^{ij} \mathbf{e}_i \otimes \mathbf{e}_j = T^i_j \mathbf{e}^i \otimes \mathbf{e}_j = T_{ij} \mathbf{e}^i \otimes \mathbf{e}^j,$$

where  $T^{ij}$ ,  $T^i_j$ , and  $T_{ij}$  are called the **contravariant**, **mixed**, and **covariant components** of  $\mathbf{T}$ , respectively. It is important to remember that these three sets of quantities form the components of the same tensor  $\mathbf{T}$  but refer to different tensor bases made up from the basis vectors of the coordinate system. Again, if we use Cartesian coordinates, the three sets of components are identical.

We may generalize the above equation to higher-order components. An object  $T^{\alpha \dots \beta}_{\gamma \dots \delta}$  is called a component of type  $(n, m)$  in which the integers  $n$  and  $m$  represent the numbers of superscripts and subscripts, respectively. By definition, components carrying only superscripts (i.e.,  $m = 0$ ) or those carrying only subscripts (i.e.,  $n = 0$ ) are referred to as the contravariant and covariant components, respectively; all others are called mixed components.

*Remark.* The order of indices needs caution. For instance, we shall see later that in general

$$T^i_j \neq T_j^i.$$

Nevertheless, we can write  $T^i_j$  with no clarification of the order of  $i$  and  $j$  if no ambiguity occurs or the order of indices is irrelevant.

### 19.1.8 Kronecker Delta

The **Kronecker delta** is a special kind of a second-order tensor that has mixed components given by  $\delta^i_j$ , and is defined as follows:

$$\delta^i_j = \begin{cases} 1 & (i = j), \\ 0 & (i \neq j). \end{cases}$$

As these are mixed components of a tensor, they transform as

$$\delta'^i_j = \frac{\partial u'^i}{\partial u^\ell} \frac{\partial u^\ell}{\partial u'^j} \delta^\ell_m = \frac{\partial u'^i}{\partial u^\ell} \frac{\partial u^\ell}{\partial u'^j} = \frac{\partial u'^i}{\partial u'^j} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{if } i \neq j, \end{cases}$$

since in the last partial derivative,  $u'^i$  and  $u'^j$  are independent coordinates. Thus, we obtain the result

$$\delta'^i_j = \delta^i_j, \quad (19.16)$$

which means that the tensor consisting of  $\delta^i_j$  has the same components in all coordinate systems. This is why the tensor consisting of  $\delta^i_j$  is called the **fundamental mixed tensor**.

*Remark.* The components  $\delta^{ij}$  (or  $\delta_{ij}$ ) are of no special importance, since they do not satisfy the invariance condition (19.16), which means that their values change when we use other coordinate systems. An exception is the case of rectangular coordinate systems, where the contravariant and covariant tensors become identical, so that we have  $\delta^{ij} = \delta^i_j = \delta_{ij}$ .

## 19.2 Metric Tensor

### 19.2.1 Definition

We now introduce important quantities that describe the geometric character of the space arithmetized by a certain curvilinear coordinate system. We know that the scalar product of a vector  $\mathbf{a}$  and local basis vectors  $\mathbf{e}_i$  and  $\mathbf{e}^j$  yields

$$a_j = \mathbf{a} \cdot \mathbf{e}_j = a^i (\mathbf{e}_i \cdot \mathbf{e}_j) \quad \text{and} \quad a^j = \mathbf{a} \cdot \mathbf{e}^j = a_i (\mathbf{e}^i \cdot \mathbf{e}^j). \quad (19.17)$$

Now we introduce the following notation:

$$\mathbf{e}_i \cdot \mathbf{e}_j = g_{ij} = g_{ji}$$

and

$$\mathbf{e}^i \cdot \mathbf{e}^j = g^{ij} = g^{ji}.$$

We can then write (19.17) in the form

$$a_j = g_{jk} a^k \quad \text{and} \quad a^j = g^{jk} a_k.$$

These equations express the covariant components of the vector  $\mathbf{a}$  in terms of its contravariant components, and vice versa. We shall see that the nine quantities  $g_{ik}$  form a second-order tensor called a **metric tensor**.

♠ **Metric tensor:**

Two-index quantities defined by

$$g_{ij} = \mathbf{e}_i \cdot \mathbf{e}_j \quad \text{and} \quad g^{k\ell} = \mathbf{e}^k \cdot \mathbf{e}^\ell \quad (19.18)$$

serve as covariant and contravariant components of a second-order tensor called a **metric tensor**.

The proof of the tensor character for the above is given in Exercise 1.

*Remark.*

1. Since both  $\mathbf{e}_i$  and  $\mathbf{e}^j$  are functions of the coordinates, so are the quantities  $g_{ij}$  and  $g^{ij}$ .
2. The mixed components  $g_j^i$  of the metric tensor are identical to those of  $\delta_j^i$  since, by definition, we have

$$g_j^i = \mathbf{e}^i \cdot \mathbf{e}_j = \delta_j^i.$$

*Examples* We calculate the elements  $g_{ij}$  for cylindrical coordinates, where  $(u^1, u^2, u^3) = (\rho, \phi, z)$  and  $\rho$  and  $\phi$  are related to Cartesian coordinates  $x$  and  $y$  as  $x = \rho \cos \phi$  and  $y = \rho \sin \phi$ . Hence, the position vector  $\mathbf{r}$  of any point may be written as

$$\mathbf{r} = \rho \cos \phi \mathbf{i} + \rho \sin \phi \mathbf{j} + z \mathbf{k},$$

where  $\mathbf{i}, \mathbf{j}, \mathbf{k}$  are orthogonal basis vectors. By definition, we have

$$\mathbf{e}_i = \frac{\partial \mathbf{r}}{\partial \rho} = \cos \phi \mathbf{i} + \sin \phi \mathbf{j}, \quad (19.19)$$

$$\mathbf{e}_j = \frac{\partial \mathbf{r}}{\partial \phi} = -\rho \sin \phi \mathbf{i} + \rho \cos \phi \mathbf{j}, \quad (19.20)$$

$$\mathbf{e}_k = \frac{\partial \mathbf{r}}{\partial z} = \mathbf{k}. \quad (19.21)$$

Thus the components of the metric tensor  $[g_{ij}] = [\mathbf{e}_i \cdot \mathbf{e}_j]$  are found to be

$$[g_{ij}] = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \rho^2 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

### 19.2.2 Geometric Role of Metric Tensors

The quantities  $g_{ik}$  (or  $g^{ik}$ ) describe the fundamental geometric character of a space arithmetized by a certain  $u^i$ -coordinate system with a **basis**  $\{\mathbf{e}_i\}$ .

A geometric role for  $g_{ij}$  was implied in the definition (19.18), where  $g_{ij}$  equals the scalar product of the two covariant local basis vectors  $\mathbf{e}_i$  and  $\mathbf{e}_j$ . Hence,  $g_{ij}$  determines the angles of local basis vectors  $\mathbf{e}_i$  and  $\mathbf{e}_j$  at each point and thus describes the coordinate( $u^k$ )-dependence of the vectors  $\mathbf{e}_i = \mathbf{e}_i(u^k)$  and  $\mathbf{e}_j = \mathbf{e}_j(u^k)$  that span the space being considered. This implies the possibility that the metric tensor  $\mathbf{g}$  rather than the basis vectors can be regarded as a more fundamental object determining the geometric nature of the space in question. Indeed, we can establish the framework of tensor calculus based on a knowledge of the spatial dependence of the metric tensor  $\mathbf{g}$  without any information about the local basis vectors. This point is dealt with in §20.3.5 Sect. 20.3.5.

The role of  $g_{ij}$  in determining the geometric nature of the space also follows from another stand point as shown below. Let  $ds$  be the arc length between two infinitely close points. We denote by  $d\mathbf{r}$  the vector joining the two points, whose covariant components are  $du_i$  and contravariant components  $du^i$ . Then, since  $d\mathbf{r} = \mathbf{e}_i du^i = \mathbf{e}^k du_k$ , we have

$$\begin{aligned}(ds)^2 &= |d\mathbf{r}|^2 = d\mathbf{r} \cdot d\mathbf{r} \\ &= \mathbf{e}_i du^i \cdot \mathbf{e}_k du^k = \mathbf{e}_i du^i \cdot \mathbf{e}^k du_k = \mathbf{e}^i du_i \cdot \mathbf{e}^k du_k,\end{aligned}$$

or

$$(ds)^2 = g_{ik} du^i du^k, \quad (19.22)$$

$$(ds)^2 = g^{ik} du_i du_k, \quad (19.23)$$

$$(ds)^2 = du_i du^i. \quad (19.24)$$

Since  $(ds)^2$  is a scalar, all of the quantities on the right-hand sides are also scalars. It should also be noted that in (19.22) and (19.23), the  $du^i$  (or  $du_k$ ) are contravariant (or covariant) components of a vector. Hence, in view of the quotient theorem regarding two-index quantities (see Sect 18.4.7), it turns out that the symmetric quantities  $g_{ik}$  (or  $g^{ik}$ ) form covariant (or contravariant) components of a second-order tensor.

### 19.2.3 Riemann Space and Metric Tensor

We have seen that in terms of tensor calculus, the metric tensor  $\mathbf{g}$  rather than the local basis vectors  $\mathbf{e}_i$  and  $\mathbf{e}^j$  is a more fundamental object in determining geometric properties of the space being considered. In fact, an abstract space of points to which we assign a certain class of a second-order tensor  $\mathbf{g}$  at each point is referred to by a special name as stated below, which gives a formal definition of the metric tensor  $\mathbf{g}$  in the language of tensor calculus.

♠ **Riemann space:**

A finite-dimensional space of points labeled by an ordered set of real coordinates  $u^1, u^2, \dots, u^n$  is called a **Riemann space** if it is possible to define two-index quantities  $g_{ij}$  that possess the following properties:

1. Each entity  $g_{ij}(u^1, u^2, \dots, u^n)$  is a real single-valued function of the coordinates and has continuous partial derivatives.
2.  $g_{ik}(u^1, u^2, \dots, u^n) = g_{ki}(u^1, u^2, \dots, u^n)$ .
3.  $g = \det[g_{ik}] \neq 0$ .

The tensor  $\mathbf{g}$  formed by the two-index quantities  $g_{ij}$  noted above is called a **metric tensor** of the space.

**Remark.** Note that the above definition of a metric tensor is free of the concept of local basis vectors.

In this context, the superscripted components  $g^{ij}$  are defined by

$$g^{ik} g_{kj} = \delta_j^i \quad \text{or} \quad g^{ik} = \frac{C^{ik}}{g},$$

where  $C^{ik}(=C^{ki})$  is the cofactor of  $g_{ik}$  in the determinant  $g = \det[g_{ik}]$ . (See Exercise 2 for the proof of the above.)

Our familiar Euclidean space is a particular class of Riemann space as stated below.

♠ **Flat Riemann space:**

A Riemann space is **flat** if and only if it admits a system of rectangular Cartesian coordinates  $x^1, x^2, \dots, x^n$  such that at every point of the space,

$$(ds)^2 = \varepsilon_1 (dx^1)^2 + \varepsilon_2 (dx^2)^2 + \dots + \varepsilon_n (dx^n)^2, \quad (19.25)$$

where each  $\varepsilon_i$  equals either +1 or -1.

♠ **Euclidean space:**

A **Euclidean space** is a flat Riemann space for which all  $\varepsilon_i$  in (19.25) are equal to +1.

#### 19.2.4 Elements of Arc, Area, and Volume

Below we describe several useful relations in connection with the elements of arc length, areas, and volumes in terms of metric tensors.

**1. Element of arc length:**

The element of arc length  $ds_i$  along a particular coordinate curve  $u^i$  with fixed  $i$  is

$$ds_i = |d\mathbf{r}| = |\mathbf{e}_i| du^i = \sqrt{\mathbf{e}_i \cdot \mathbf{e}_i} du^i = \sqrt{g_{ii}} du^i \quad (\text{no summation over } i).$$

**2. Element of area**

The element of area  $d\sigma_1$  in the coordinate surface  $u^1 = \text{const}$ ; for instance, reads

$$\begin{aligned} d\sigma_1 &= |d\mathbf{r}_2 \times d\mathbf{r}_3| = |\mathbf{e}_2 \times \mathbf{e}_3| du^2 du^3 \\ &= \sqrt{(\mathbf{e}_2 \times \mathbf{e}_3) \cdot (\mathbf{e}_2 \times \mathbf{e}_3)} du^2 du^3 \\ &= \sqrt{(\mathbf{e}_2 \cdot \mathbf{e}_2)(\mathbf{e}_3 \cdot \mathbf{e}_3) - (\mathbf{e}_2 \cdot \mathbf{e}_3)(\mathbf{e}_2 \cdot \mathbf{e}_3)} du^2 du^3 \\ &= \sqrt{g_{22}g_{33} - (g_{23})^2} du^2 du^3. \end{aligned}$$

Similarly, we have

$$\begin{aligned} d\sigma_2 &= \sqrt{g_{33}g_{11} - (g_{13})^2} du^3 du^1, \\ d\sigma_3 &= \sqrt{g_{11}g_{22} - (g_{12})^2} du^1 du^2, \end{aligned}$$

which are summarized by

$$d\sigma_i = \sqrt{g_{jj}g_{kk} - (g_{jk})^2} du^j du^k \quad (\text{no summation over } j \text{ and } k),$$

where  $i, j, k$  is a cyclic permutation of the numbers 1, 2, 3.

**3. Element of volume**

Finally, we can derive the equation for the element of volume as

$$\begin{aligned} dV &= |(d\mathbf{r}_1 \times d\mathbf{r}_2) \cdot d\mathbf{r}_3| = |(\mathbf{e}_1 \times \mathbf{e}_2) \cdot \mathbf{e}_3| du^1 du^2 du^3 \\ &= \sqrt{g} du^1 du^2 du^3, \end{aligned}$$

where  $g = \det[g_{ik}]$ . [Proof of the identity  $(\mathbf{e}_1 \times \mathbf{e}_2) \cdot \mathbf{e}_3 = g$  is given in Exercise 2.]

Our results are summarized as:

**♠ Theorem:**

Elements of arc length  $ds_i$ , area  $d\sigma_i$ , and volume  $dV$ , respectively, are represented in terms of curvilinear coordinate systems by

$$ds_i = \sqrt{g_{ii}} du^i \quad (\text{no sum over } i),$$

$$d\sigma_i = \sqrt{g_{jj}g_{kk} - (g_{jk})^2} du^j du^k \quad (\text{no sum over } j \text{ and } k), \text{ and}$$

$$dV = \sqrt{g} du^1 du^2 du^3,$$

where  $i, j, k$  is a cyclic permutation of 1, 2, 3.

### 19.2.5 Scale Factors

In this subsection, we consider the case of orthogonal coordinate systems, for which the basic descriptive quantities are the **scale factors** (or the **metric coefficients**)  $h_1, h_2, h_3$ , defined by

$$h_1 = \sqrt{g_{11}}, \quad h_2 = \sqrt{g_{22}}, \quad h_3 = \sqrt{g_{33}}.$$

Obviously, they satisfy the equation

$$(ds)^2 = (h_1 du^1)^2 + (h_2 du^2)^2 + (h_3 du^3)^2.$$

Furthermore, since  $g_{ij} = 0$  for  $i \neq j$ , we have

$$ds_i = h_i du^i \quad (\text{no sum over } i),$$

$$d\sigma_i = h_j h_k du^j du^k \quad (\text{no sum over } j \text{ and } k),$$

$$dV = h_1 h_2 h_3 du^1 du^2 du^3,$$

where  $i, j, k$  is a cyclic permutation of 1, 2, 3.

*Examples* **1.** In rectangular Cartesian coordinates,

$$(ds)^2 = (dx)^2 + (dy)^2 + (dz)^2,$$

so

$$h_1 = h_2 = h_3 = 1.$$

**2.** In cylindrical coordinates,

$$(ds)^2 = (dR)^2 + (Rd\theta)^2 + (dz)^2,$$

so

$$h_1 = 1, \quad h_2 = R, \quad h_3 = 1.$$

**3.** In spherical coordinates,

$$(ds)^2 = (dR)^2 + (Rd\theta)^2 + (R \sin \theta d\phi)^2,$$

so

$$h_1 = 1, \quad h_2 = R, \quad h_3 = R \sin \theta.$$

### 19.2.6 Representation of Basis Vectors in Derivatives

It is often desirable to represent local covariant basis vectors  $\mathbf{e}_i$  as well as components of metric tensors  $g_{ij} = \mathbf{e}_i \cdot \mathbf{e}_j$  at a point  $\mathbf{r}$  in terms of derivatives of  $\mathbf{r}$  with respect to coordinates the  $u^i$ .

Suppose the relation between a system of curvilinear coordinates  $u^1, u^2, u^3$  and an underlying system of rectangular coordinates  $x_1, x_2, x_3 (= x, y, z)$  is given by

$$u^i = u^i(x_k) \quad \text{and} \quad x_k = x_k(u^i), \quad (19.26)$$

where the Jacobian

$$J = \left| \frac{\partial u^i}{\partial x_k} \right|$$

is neither zero nor infinite. Writing the latter equation in (19.26) more concisely as

$$\mathbf{r} = \mathbf{r}(u^i),$$

where  $\mathbf{r} = x_k \mathbf{i}_k$  is the position arrow of an arbitrary point, we find

$$d\mathbf{r} = \frac{\partial \mathbf{r}}{\partial u^i} du^i.$$

It then follows that

$$(ds)^2 = d\mathbf{r} \cdot d\mathbf{r} = \frac{\partial \mathbf{r}}{\partial u^i} \cdot \frac{\partial \mathbf{r}}{\partial u^j} du^i du^j,$$

which implies that the vectors of the local basis are

$$\mathbf{e}_i = \frac{\partial \mathbf{r}}{\partial u^i}$$

and the metric tensor is

$$g_{ij} = \frac{\partial \mathbf{r}}{\partial u^i} \cdot \frac{\partial \mathbf{r}}{\partial u^j} = \frac{\partial x_k}{\partial u^i} \mathbf{i}_k \cdot \frac{\partial x_\ell}{\partial u^j} \mathbf{i}_\ell = \frac{\partial x_k}{\partial u^i} \frac{\partial x_\ell}{\partial u^j} (\mathbf{i}_k \cdot \mathbf{i}_\ell) = \frac{\partial x_k}{\partial u^i} \frac{\partial x_k}{\partial u^j}.$$

This leads to the following expression for the scale factors (for the case of orthogonal coordinate systems):

$$h_i = \sqrt{\left( \frac{\partial x_1}{\partial u^i} \right)^2 + \left( \frac{\partial x_2}{\partial u^i} \right)^2 + \left( \frac{\partial x_3}{\partial u^i} \right)^2}.$$

### 19.2.7 Index Lowering and Raising

In curvilinear coordinate systems, it is possible to express a scalar product of two vectors via several different subscript forms. For instance, the scalar product of two vectors  $\mathbf{a}$  and  $\mathbf{b}$  may be written using their contravariant or covariant components:

$$\mathbf{a} \cdot \mathbf{b} = a^i \mathbf{e}_i \cdot b^j \mathbf{e}_j = g_{ij} a^i b^j \quad (19.27)$$

and

$$\mathbf{a} \cdot \mathbf{b} = a_i \mathbf{e}^i \cdot b_j \mathbf{e}^j = g^{ij} a_i b_j. \quad (19.28)$$

Furthermore, we may express the scalar product in terms of the contravariant components of one vector and the covariant components of the other:

$$\mathbf{a} \cdot \mathbf{b} = a_i \mathbf{e}^i \cdot b^j \mathbf{e}_j = a_i b^j \delta_j^i = a_i b^i \quad (19.29)$$

and

$$\mathbf{a} \cdot \mathbf{b} = a^i \mathbf{e}_i \cdot b_j \mathbf{e}^j = a^i b_j \delta_i^j = a^i b_i. \quad (19.30)$$

By comparing the four alternative expressions (19.27)–(19.30) for  $\mathbf{a} \cdot \mathbf{b}$ , we can deduce the following useful property of  $g_{ij}$  and  $g^{ij}$ . From (19.27) and (19.30) we see that the identity

$$g_{ij} a^i b^j = a^i b_i$$

holds for any arbitrary  $a^i$ . Hence, we have

$$g_{ij} b^j = b_i. \quad (19.31)$$

which illustrates the fact that the covariant components  $g_{ij}$  can be used to lower an index  $b^j$ . In other words, it provides a means of obtaining the covariant components  $b_i$  of a vector from its contravariant components  $b^j$ . By a similar argument, we have

$$g^{ij} b_j = b^i,$$

where the contravariant components  $g^{ij}$  are used to raise the index  $j$  attached to  $b_j$ .

#### ♠ Index lowering and raising (I):

For any vector  $\mathbf{a}$ , its components  $a^i$  and  $a_i$  are related via the components of the metric tensor as

$$a_i = g_{ik} a^k \quad \text{and} \quad a^j = g^{j\ell} a_\ell.$$

The above discussion regarding vectors can be extended to tensors of arbitrary rank. For example, the contraction with  $g_{ij}$  results in a lowering of the corresponding index:

$$T_{ij} = g_{ik} T_{\bullet j}^k = g_{jk} T_i^{\bullet k}. \quad (19.32)$$

Here the dots ( $\bullet$ ) in the mixed components emphasize the order of occurrence of the indices; in fact, in general,  $T_i^{\bullet l} \neq T_{\bullet i}^l$ . Repeated contraction with  $g_{ij}$  yields

$$T_{ij} = g_{ik} g_{jl} T^{kl}.$$

Similarly, contraction with  $g^{ij}$  raises an index, i.e.,

$$T^{ij} = g^{ik} T_k^{\bullet j} = g^{ik} g^{jl} T_{kl}. \quad (19.33)$$

Comparable arguments are applicable to local basis vectors  $\mathbf{e}_i$  and  $\mathbf{e}^j$  as stated below.

♠ **Index lowering and raising (II):**

Local basis vectors  $\mathbf{e}_i$  and  $\mathbf{e}^k$  are related as

$$\mathbf{e}_i = g_{ik} \mathbf{e}^k \quad \text{and} \quad \mathbf{e}^j = g^{j\ell} \mathbf{e}_\ell.$$

*Proof* Since  $\mathbf{a} = a^i \mathbf{e}_i = a_j \mathbf{e}^j = a^k g_{kj} \mathbf{e}^j$ , we have

$$a^1 (\mathbf{e}_1 - g_{1j} \mathbf{e}^j) + a^2 (\mathbf{e}_2 - g_{2j} \mathbf{e}^j) + a^3 (\mathbf{e}_3 - g_{3j} \mathbf{e}^j) = \mathbf{0},$$

which holds for any vector  $\mathbf{a}$ . Hence,  $\mathbf{e}_k - g_{kj} \mathbf{e}^j = \mathbf{0}$  for all  $k$ , i.e.,

$$\mathbf{e}_k = g_{kj} \mathbf{e}^j.$$

Similarly, we have

$$\mathbf{e}^k = g^{kj} \mathbf{e}_j. \quad \clubsuit$$

## Exercises

1. Show that the quantities  $g_{ij} = \mathbf{e}_i \cdot \mathbf{e}_j$  form the covariant components of a second-order tensor.

**Solution:** In the new (primed) coordinate system we have  $g'_{ij} = \mathbf{e}'_i \cdot \mathbf{e}'_j$ . Using the transformation law (19.7) of covariant basis vectors, we have

$$g'_{ij} = \left( \frac{\partial u^k}{\partial u'^i} \mathbf{e}_k \right) \cdot \left( \frac{\partial u^l}{\partial u'^j} \mathbf{e}_l \right) = \frac{\partial u^k}{\partial u'^i} \frac{\partial u^l}{\partial u'^j} (\mathbf{e}_k \cdot \mathbf{e}_l) = \frac{\partial u^k}{\partial u'^i} \frac{\partial u^l}{\partial u'^j} g_{kl}.$$

This clearly indicates that the  $g_{ij}$  are covariant components of a second-order tensor (i.e., the metric tensor  $\mathbf{g}$ ). A similar argument shows that the quantities  $g^{ij}$  form the contravariant components of  $\mathbf{g}$ , which transform as follows:

$$g'^{ij} = \frac{\partial u'^i}{\partial u^k} \frac{\partial u'^j}{\partial u^l} g^{kl}. \quad \clubsuit$$

2. Show that the matrix  $[g^{ij}]$  is the inverse of the matrix  $[g_{ij}]$ .

For an arbitrary vector  $\mathbf{a}$ , we find  $a^i = g^{ij}a_j = g^{ij}g_{jk}a^k$ . Since  $\mathbf{a}$  is arbitrary, we must have

$$g^{ij}g_{jk} = \delta_k^i = \begin{cases} 1 & i = k, \\ 0 & i \neq k. \end{cases} \quad (19.34)$$

This clearly indicates that the matrices  $[g_{ij}]$  and  $[g^{ij}]$  are inverse to each other. ♣

3. Show that  $\sqrt{g} = (\mathbf{e}_i \times \mathbf{e}_j) \cdot \mathbf{e}_k$  and  $1/\sqrt{g} = (\mathbf{e}^i \times \mathbf{e}^j) \cdot \mathbf{e}^k$ , where  $i, j, k$  is a cyclic permutation of the numbers 1, 2, 3.

**Solution:** By direct calculations, we obtain

$$g^{i\ell} = \mathbf{e}^i \cdot \mathbf{e}^\ell = \frac{(\mathbf{e}_j \times \mathbf{e}_k) \cdot (\mathbf{e}_m \times \mathbf{e}_n)}{[\mathbf{e}_i \cdot (\mathbf{e}_j \times \mathbf{e}_k)][\mathbf{e}_\ell \cdot (\mathbf{e}_m \times \mathbf{e}_n)]}, \quad (19.35)$$

where  $i, j, k$  and  $\ell, m, n$  are cyclic permutations of the ordered set of numbers 1, 2, 3. The numerator in (19.35) reads

$$\begin{aligned} (\mathbf{e}_j \times \mathbf{e}_k) \cdot (\mathbf{e}_m \times \mathbf{e}_n) &= [(\mathbf{e}_j \times \mathbf{e}_k) \times \mathbf{e}_m] \cdot \mathbf{e}_n \\ &= [(\mathbf{e}_j \cdot \mathbf{e}_m) \mathbf{e}_k - (\mathbf{e}_k \cdot \mathbf{e}_m) \mathbf{e}_j] \cdot \mathbf{e}_n \\ &= (\mathbf{e}_j \cdot \mathbf{e}_m) (\mathbf{e}_k \cdot \mathbf{e}_n) - (\mathbf{e}_k \cdot \mathbf{e}_m) (\mathbf{e}_j \cdot \mathbf{e}_n) \\ &= \begin{vmatrix} \mathbf{e}_j \cdot \mathbf{e}_m & \mathbf{e}_k \cdot \mathbf{e}_m \\ \mathbf{e}_j \cdot \mathbf{e}_n & \mathbf{e}_k \cdot \mathbf{e}_n \end{vmatrix} = \begin{vmatrix} g_{jm} & g_{km} \\ g_{jn} & g_{kn} \end{vmatrix} = C^{i\ell}. \end{aligned}$$

Here,  $C^{i\ell}$  is the cofactor of  $g_{i\ell}$  in the determinant  $g = \det[g_{i\ell}]$ . Comparing the results with the definition  $g^{i\ell} = C^{i\ell}/g$ , we find that

$$g = [\mathbf{e}_i \cdot (\mathbf{e}_j \times \mathbf{e}_k)][\mathbf{e}_\ell \cdot (\mathbf{e}_m \times \mathbf{e}_n)],$$

which is equivalent to

$$g = [\mathbf{e}_i \cdot (\mathbf{e}_j \times \mathbf{e}_k)]^2, \quad \text{i.e.,} \quad \sqrt{g} = \pm \mathbf{e}_i \cdot (\mathbf{e}_j \times \mathbf{e}_k),$$

where the plus sign is chosen if the given basis is right-handed. In a similar manner, via the relations  $g_{i\ell} = C_{i\ell}$  and  $\det[\delta_i^k] = \det[g_{ij}g^{jk}] = \det[g_{ij}]\det[g^{jk}] = 1$ , we obtain

$$\frac{1}{\sqrt{g}} = \pm \mathbf{e}^i \cdot (\mathbf{e}^j \times \mathbf{e}^k). \quad \clubsuit$$

## 19.3 Christoffel Symbols

### 19.3.1 Derivatives of Basis Vectors

Several new concepts are required for the differentiation of vectors or tensors with respect to curvilinear coordinates. Recall that in a general curvilinear coordinate system, the basis vectors  $\mathbf{e}_i$  and  $\mathbf{e}^i$  are functions of the coordinates. This implies that differentiation of vectors (say,  $\mathbf{v} = v^i \mathbf{e}_i$ ) or tensors (say,  $\mathbf{T} = T_i^j \mathbf{e}^i \otimes \mathbf{e}_j$ ) involves their derivatives, such as  $\partial \mathbf{e}_i / \partial u^j$ .

Suppose that the derivative  $\partial \mathbf{e}_i / \partial u^j$  can be written as a linear combination of the basis vectors  $\mathbf{e}_k$  as denoted by

$$\frac{\partial \mathbf{e}_i}{\partial u^j} = \Gamma_{ij}^k \mathbf{e}_k, \quad (19.36)$$

the symbol  $\Gamma_{ij}^k$  being the coefficients associated with the  $k$ th component of the linear combination. Using the reciprocity relation  $\mathbf{e}^i \cdot \mathbf{e}_j = \delta_j^i$ , we write this as

$$\Gamma_{ij}^k = \mathbf{e}^k \cdot \frac{\partial \mathbf{e}_i}{\partial u^j}. \quad (19.37)$$

This three-index symbol is called a **Christoffel symbol**. In a similar manner as above, we can show that the derivative of the contravariant basis vectors reads

$$\frac{\partial \mathbf{e}^i}{\partial u^j} = -\Gamma_{kj}^i \mathbf{e}^k. \quad (19.38)$$

Details of the derivation are given in Exercise 1.

We shall see that Christoffel symbols play a key role in defining the derivatives of vectors and tensors in terms of general coordinate systems. A more formal definition of Christoffel symbols in terms of metric tensors is given in Sect. 19.3.4.

**Remark.** It is clear from (19.37) that in Cartesian coordinate systems,  $\Gamma_{ij}^k = 0$  for all values of the indices  $i, j$ , and  $k$ , owing to the identity:  $\partial \mathbf{e}_i / \partial u^j \equiv 0$ .

*Example 4.* Let us calculate the Christoffel symbols  $\Gamma_{ij}^m$  for cylindrical coordinates, where  $(u^1, u^2, u^3) = (\rho, \phi, z)$ , and the position vector  $\mathbf{r}$  of any point may be written

$$\mathbf{r} = \rho \cos \phi \mathbf{i} + \rho \sin \phi \mathbf{j} + z \mathbf{k}.$$

From this we find that the covariant basis vectors are given by

$$\mathbf{e}_\rho = \frac{\partial \mathbf{r}}{\partial \rho} = \cos \phi \mathbf{i} + \sin \phi \mathbf{j}, \quad (19.39)$$

$$\mathbf{e}_\phi = \frac{\partial \mathbf{r}}{\partial \phi} = -\rho \sin \phi \mathbf{i} + \rho \cos \phi \mathbf{j}, \quad (19.40)$$

$$\mathbf{e}_z = \frac{\partial \mathbf{r}}{\partial z} = \mathbf{k}. \quad (19.41)$$

It is a straightforward matter to show that the only derivatives of these vectors that are nonzero with respect to the coordinates are

$$\frac{\partial \mathbf{e}_\rho}{\partial \phi} = \frac{1}{\rho} \mathbf{e}_\phi, \quad \frac{\partial \mathbf{e}_\phi}{\partial \rho} = \frac{1}{\rho} \mathbf{e}_\phi, \quad \frac{\partial \mathbf{e}_\phi}{\partial \phi} = -\rho \mathbf{e}_\rho.$$

Thus, from (19.37), we immediately have

$$\Gamma_{12}^2 = \Gamma_{21}^2 = \frac{1}{\rho} \quad \text{and} \quad \Gamma_{22}^1 = -\rho. \quad (19.42)$$

### 19.3.2 Nontensor Character

Despite their appearance, the Christoffel symbols  $\Gamma_{ij}^k$  do not form the components of a third-order tensor.

♠ **Theorem:**

Christoffel symbols  $\Gamma_{ij}^k$  do not form any kind of tensor.

*Proof* This is verified by considering their transformation behavior under a general coordinate transformation. In a transformed coordinate system, we have

$$\Gamma'^k_{ij} = \mathbf{e}'^k \cdot \frac{\partial \mathbf{e}'_i}{\partial u'^j}. \quad (19.43)$$

Applying the transformation law of local basis vectors, we obtain

$$\begin{aligned} \Gamma'^k_{ij} &= \left( \frac{\partial u'^k}{\partial u^n} \mathbf{e}^n \right) \cdot \frac{\partial}{\partial u'^j} \left( \frac{\partial u^l}{\partial u'^i} \mathbf{e}_l \right) \\ &= \left( \frac{\partial u'^k}{\partial u^n} \mathbf{e}^n \right) \cdot \left[ \frac{\partial^2 u^l}{\partial u'^i \partial u'^j} \mathbf{e}_l + \left( \frac{\partial u^l}{\partial u'^i} \frac{\partial \mathbf{e}_l}{\partial u'^j} \right) \right] \\ &= \frac{\partial u'^k}{\partial u^n} \frac{\partial^2 u^l}{\partial u'^i \partial u'^j} (\mathbf{e}^n \cdot \mathbf{e}_l) + \frac{\partial u'^k}{\partial u^n} \frac{\partial u^l}{\partial u'^i} \left( \mathbf{e}^n \cdot \frac{\partial \mathbf{e}_l}{\partial u'^j} \right) \\ &= \frac{\partial u'^k}{\partial u^n} \frac{\partial^2 u^l}{\partial u'^i \partial u'^j} \delta_l^n + \frac{\partial u'^k}{\partial u^n} \frac{\partial u^l}{\partial u'^i} \frac{\partial u^m}{\partial u'^j} \left( \mathbf{e}^n \cdot \frac{\partial \mathbf{e}_l}{\partial u^m} \right) \\ &= \frac{\partial u'^k}{\partial u^l} \frac{\partial^2 u^l}{\partial u'^i \partial u'^j} + \frac{\partial u'^k}{\partial u^n} \frac{\partial u^l}{\partial u'^i} \frac{\partial u^m}{\partial u'^j} \Gamma_{lm}^n. \end{aligned} \quad (19.44)$$

Hence, the presence of the first term in the last line in (19.44) prevents the  $\Gamma_{ij}^k$  from forming a third-order tensor. ♣

### 19.3.3 Properties of Christoffel Symbols

Christoffel symbols  $\Gamma_{ij}^k$  satisfy the following relations:

1.  $\Gamma_{ij}^k = \Gamma_{ji}^k$ .
2.  $\frac{\partial g_{ij}}{\partial u^k} = g_{\ell j} \Gamma_{ik}^\ell + g_{i\ell} \Gamma_{jk}^\ell$ .
3.  $\frac{\partial g^{ij}}{\partial u^k} = -g^{i\ell} \Gamma_{\ell k}^j - g_{j\ell} \Gamma_{\ell k}^i$ .
4.  $\Gamma_{ik}^k = \frac{\partial}{\partial u^i} \log \sqrt{|g|} = \frac{1}{\sqrt{|g|}} \frac{\partial \sqrt{|g|}}{\partial u^i}$ .

Proofs of these relations are given in Exercises 2–4.

*Remark.* Some textbooks refer to our three-index symbol  $\Gamma_{ij}^k$  defined by (19.37) as the **Christoffel symbol of the second kind** and use the following notation:

$$\Gamma_{ij}^k = \left\{ \begin{matrix} k \\ i \ j \end{matrix} \right\} = \mathbf{e}_k \cdot \frac{\partial \mathbf{e}_i}{\partial u^j}. \quad (19.45)$$

As a counterpart, we may define the **Christoffel symbol of the first kind**  $[k, ij]$  by

$$[k, ij] = \mathbf{e}^k \cdot \frac{\partial \mathbf{e}_i}{\partial u^j}. \quad (19.46)$$

Note that the index  $k$  on the right-hand side of (19.46) is a *superscript*, whereas that of (19.45) is a *subscript*. These two kinds of Christoffel symbols are related to each other as

$$\left\{ \begin{matrix} k \\ i \ j \end{matrix} \right\} = g^{k\ell} [\ell, ij].$$

### 19.3.4 Alternative Expression

In principle, we can calculate the  $\Gamma_{ij}^k$  in a given coordinate system using the expression (19.37) based on  $\mathbf{e}_i$ . However, it is simple to use an alternative expression in terms of the metric tensor  $g_{ij}$  and its derivatives as stated below.

♠ **Theorem:**

Christoffel symbols are expressed as

$$\Gamma_{ij}^m = \frac{1}{2} g^{mk} \left( \frac{\partial g_{jk}}{\partial u^i} + \frac{\partial g_{ki}}{\partial u^j} - \frac{\partial g_{ij}}{\partial u^k} \right).$$

*Proof* Recall the relation

$$\frac{\partial g_{ij}}{\partial u^k} = g_{\ell j} \Gamma_{ik}^\ell + g_{i\ell} \Gamma_{jk}^\ell \quad (19.47)$$

given in Sect. 19.3.3. By cyclically permuting the free indices  $i, j, k$ , we obtain two further equivalent relations:

$$\frac{\partial g_{jk}}{\partial u^i} = \Gamma_{ji}^\ell g_{\ell k} + \Gamma_{ki}^\ell g_{j\ell} \quad (19.48)$$

and

$$\frac{\partial g_{ki}}{\partial u^j} = \Gamma_{kj}^\ell g_{\ell i} + \Gamma_{ij}^\ell g_{k\ell}. \quad (19.49)$$

Then, subtracting (19.47) from the sum of (19.48) and (19.49), we find

$$\begin{aligned} & \frac{\partial g_{jk}}{\partial u^i} + \frac{\partial g_{ki}}{\partial u^j} - \frac{\partial g_{ij}}{\partial u^k} \\ &= \Gamma_{ji}^\ell g_{\ell k} + \Gamma_{ki}^\ell g_{j\ell} + \Gamma_{kj}^\ell g_{\ell i} + \Gamma_{ij}^\ell g_{k\ell} - \Gamma_{ik}^\ell g_{\ell j} - \Gamma_{jk}^\ell g_{i\ell} \\ &= (\Gamma_{ji}^\ell g_{\ell k} + \Gamma_{ij}^\ell g_{k\ell}) + (\Gamma_{ki}^\ell g_{j\ell} - \Gamma_{ik}^\ell g_{\ell j}) + (\Gamma_{kj}^\ell g_{\ell i} - \Gamma_{jk}^\ell g_{i\ell}) \\ &= 2\Gamma_{ij}^\ell g_{k\ell} + 0 + 0 = 2\Gamma_{ij}^\ell g_{k\ell}, \end{aligned} \quad (19.50)$$

where we have used the symmetry properties:  $g_{ij} = g_{ji}$  and  $\Gamma_{ij}^\ell = \Gamma_{ji}^\ell$ . Contracting both sides with  $g^{mk}$  yields

$$g^{mk} \left( \frac{\partial g_{jk}}{\partial u^i} + \frac{\partial g_{ki}}{\partial u^j} - \frac{\partial g_{ij}}{\partial u^k} \right) = 2g^{mk} \Gamma_{ij}^\ell g_{k\ell} = 2\Gamma_{ij}^\ell \delta_\ell^m = 2\Gamma_{ij}^m,$$

i.e.,

$$\Gamma_{ij}^m = \frac{1}{2} g^{mk} \left( \frac{\partial g_{jk}}{\partial u^i} + \frac{\partial g_{ki}}{\partial u^j} - \frac{\partial g_{ij}}{\partial u^k} \right). \quad \clubsuit \quad (19.51)$$

This result enables us to compute the Christoffel symbol of a given coordinate system from information about the metric tensor.

*Examples* We again evaluate the Christoffel symbols  $\Gamma_{ij}^m$  for cylindrical coordinates. Using (19.51) and the fact that  $g_{11} = 1$ ,  $g_{22} = \rho^2$ ,  $g_{33} = 1$  and the other components are zero, we see that the only three nonzero Christoffel symbols are indeed  $\Gamma_{12}^2 = \Gamma_{21}^2$  and  $\Gamma_{22}^1$ . Given by

$$\Gamma_{12}^2 = \Gamma_{21}^2 = \frac{1}{2g_{22}} \frac{\partial g_{22}}{\partial u^1} = \frac{1}{2\rho^2} \frac{\partial}{\partial \rho} \rho^2 = \frac{1}{\rho}, \quad (19.52)$$

$$\Gamma_{22}^1 = -\frac{1}{2g_{11}} \frac{\partial g_{22}}{\partial u^1} = -\frac{1}{2} \frac{\partial}{\partial \rho} \rho^2 = -\rho, \quad (19.53)$$

they agree with the expressions in (19.42).

**Remark.** The result (19.51) implies that the Christoffel symbol of the first kind  $[k, ij]$  mentioned in (19.46) is written as

$$[k, ij] = \frac{1}{2} \left( \frac{\partial g_{jk}}{\partial u^i} + \frac{\partial g_{ki}}{\partial u^j} - \frac{\partial g_{ij}}{\partial u^k} \right).$$

## Exercises

1. Derive equation (19.38).

**Solution:** By differentiating the reciprocity relation  $\mathbf{e}^i \cdot \mathbf{e}_j = \delta_j^i$  with respect to the coordinates, we have

$$\frac{\partial \mathbf{e}^i}{\partial u^k} \cdot \mathbf{e}_j + \mathbf{e}^i \cdot \frac{\partial \mathbf{e}_j}{\partial u^k} = \frac{\partial \delta_j^i}{\partial u^k}. \quad (19.54)$$

The right-hand side of (19.54) vanishes since the element  $\delta_j^i$  consists of the constants +1 and 0, which are independent of the coordinates  $u^k$ . Hence, using (19.37), we obtain

$$\frac{\partial \mathbf{e}^i}{\partial u^k} \cdot \mathbf{e}_j + \Gamma_{jk}^i = 0. \quad (19.55)$$

Similar to the case of (19.36), for the moment we write the derivative  $\partial \mathbf{e}^i / \partial u^j$  as a linear combination of the basis vector  $\mathbf{e}^\ell$  as

$$\frac{\partial \mathbf{e}^i}{\partial u^k} = B_{ik}^\ell \mathbf{e}^\ell. \quad (19.56)$$

Substituting (19.56) into (19.55), we obtain  $B_{ik}^j = -\Gamma_{jk}^i$ . Consequently, we have  $\frac{\partial \mathbf{e}^i}{\partial u^k} = -\Gamma_{\ell k}^i \mathbf{e}^\ell$ , or equivalently (by interchanging the subscripts),

$$\frac{\partial \mathbf{e}^i}{\partial u^j} = -\Gamma_{kj}^i \mathbf{e}^k. \quad \clubsuit$$

2. Show that  $\Gamma_{ij}^k = \Gamma_{ji}^k$ .

**Solution:** It follows that  $\frac{\partial \mathbf{e}_i}{\partial u^j} = \frac{\partial}{\partial u^j} \frac{\partial \mathbf{r}}{\partial u^i} = \frac{\partial}{\partial u^i} \frac{\partial \mathbf{r}}{\partial u^j} = \frac{\partial \mathbf{e}_j}{\partial u^i}$ , which yields  $\Gamma_{ij}^k = \mathbf{e}^k \cdot \frac{\partial \mathbf{e}_i}{\partial u^j} = \mathbf{e}^k \cdot \frac{\partial \mathbf{e}_j}{\partial u^i} = \Gamma_{ji}^k$ .  $\clubsuit$

3. Show that  $\frac{\partial g_{ij}}{\partial u^k} = g_{\ell j} \Gamma_{ik}^\ell + g_{i\ell} \Gamma_{jk}^\ell$ .

**Solution:** Derivatives of the metric tensor  $g_{ij} = \mathbf{e}_i \cdot \mathbf{e}_j$  with respect to  $u^k$  read

$$\begin{aligned}\frac{\partial g_{ij}}{\partial u^k} &= \frac{\partial \mathbf{e}_i}{\partial u^k} \cdot \mathbf{e}_j + \mathbf{e}_i \cdot \frac{\partial \mathbf{e}_j}{\partial u^k} = \Gamma_{ik}^\ell \mathbf{e}_\ell \cdot \mathbf{e}_j + \mathbf{e}_i \cdot \Gamma_{jk}^\ell \mathbf{e}_\ell \\ &= \Gamma_{ik}^\ell g_{\ell j} + \Gamma_{jk}^\ell g_{i\ell}. \quad \clubsuit\end{aligned}$$

4. Show that  $\frac{\partial g}{\partial u^k} = g g^{ji} \frac{\partial g_{ij}}{\partial u^k}$ , where  $g = \det[g_{ij}]$ .

**Solution:** We know that the determinant  $g$  is given by (see Sect. 18.1.7)

$$g = \sum_{j=1}^n g_{ij} C^{ij} \quad (\text{with } i \text{ fixed}),$$

where  $C^{ij}$  is the cofactor of the element  $g_{ij}$  in  $g$ . Partially differentiating both sides with respect to  $g_{ij}$  gives

$$\frac{\partial g}{\partial g_{ij}} = C^{ij}. \quad (19.57)$$

Since  $g^{ij} = C^{ij}/g$  (see Sect. 19.2.4), it follows from (19.57) that

$$\frac{\partial g}{\partial u^k} = \frac{\partial g}{\partial g_{ij}} \frac{\partial g_{ij}}{\partial u^k} = C^{ij} \frac{\partial g_{ij}}{\partial u^k} = g g^{ji} \frac{\partial g_{ij}}{\partial u^k}. \quad \clubsuit$$

5. Show that  $\Gamma_{ik}^k = \frac{\partial}{\partial u^i} \log \sqrt{g}$ .

**Solution:** According to the expression (19.51), we have

$$\Gamma_{ki}^i = \frac{1}{2} g^{i\ell} \left( \frac{\partial g_{i\ell}}{\partial u^k} + \frac{\partial g_{k\ell}}{\partial u^i} - \frac{\partial g_{ki}}{\partial u^\ell} \right).$$

The last two terms in the parentheses cancel out because

$$g^{i\ell} \frac{\partial g_{k\ell}}{\partial u^i} = g^{\ell i} \frac{\partial g_{ki}}{\partial u^\ell} = g^{i\ell} \frac{\partial g_{ki}}{\partial u^\ell},$$

where we have interchanged the dummy indices  $i$  and  $l$  in the first equality, and have used the symmetry of the metric tensor in the second. Hence, we set

$$\Gamma_{ki}^i = \frac{g^{i\ell}}{2} \frac{\partial g_{i\ell}}{\partial u^k}. \quad (19.58)$$

This can be further simplified by using the result of Exercise 4 as

$$\Gamma_{ki}^i = \frac{1}{2g} \frac{\partial g}{\partial u^k} = \frac{1}{2g} \frac{\partial g}{\partial \sqrt{g}} \frac{\partial \sqrt{g}}{\partial u^k} = \frac{1}{\sqrt{g}} \frac{\partial \sqrt{g}}{\partial u^k} = \frac{\partial}{\partial u^k} \log \sqrt{g}. \quad \clubsuit$$

## 19.4 Covariant Derivatives

### 19.4.1 Covariant Derivatives of Vectors

The derivatives of a scalar in terms of Cartesian coordinates work as covariant components of a vector. This is also true for the case of general coordinate systems, as can be shown by considering the differential of a scalar

$$d\phi = \frac{\partial\phi}{\partial u^i} du^i.$$

Since the  $du^i$  are contravariant components of a vector and  $d\phi$  is a scalar, we see from the quotient law that the quantities  $\partial\phi/\partial u^i$  must form covariant components of a vector.

Except for a scalar, however, the derivatives of a general tensor do not necessarily form the component of another tensor. To see this, we consider the derivative of the covariant components  $v^i$  of a vector  $\mathbf{v}$  with respect to a general coordinate  $u^j$ . In a new (primed) coordinate, it reads

$$\begin{aligned} \frac{\partial v'^i}{\partial u'^j} &= \frac{\partial u^k}{\partial u'^j} \frac{\partial v'^i}{\partial u^k} = \frac{\partial u^k}{\partial u'^j} \frac{\partial}{\partial u^k} \left( \frac{\partial u'^i}{\partial u^l} v^l \right) \\ &= \frac{\partial u^k}{\partial u'^j} \frac{\partial u'^i}{\partial u^l} \frac{\partial v^l}{\partial u^k} + \frac{\partial u^k}{\partial u'^j} \frac{\partial^2 u'^i}{\partial u^k \partial u^l} v^l. \end{aligned} \quad (19.59)$$

The presence of the second term in the last line of (19.59) prevents the derivative  $\partial v^i/\partial x^j$  from obeying the transformation law of the components of a second-order tensor. The nontensor character stems from the fact that the second-order derivative,

$$\frac{\partial^2 u'^i}{\partial u^k \partial u^l}, \quad (19.60)$$

involved in the last line of (19.59) does not vanish. In fact, the first-order derivative  $\partial u'^i/\partial u^l$  is not constant in non-Cartesian coordinates, whereas it is constant in Cartesian coordinates [so that the term (19.60) vanishes in the latter case].

In the context above, it is natural to introduce a new class of differentiation that turns the derivatives of components of a tensor into components of another tensor. This is achieved with the help of the Christoffel symbols discussed in Sect. 19.3. Let us consider the derivative of a vector  $\mathbf{v}$  with respect to the coordinates  $u^j$ . We find

$$\frac{\partial \mathbf{v}}{\partial u^j} = \frac{\partial v^i}{\partial u^j} \mathbf{e}_i + v^i \frac{\partial \mathbf{e}_i}{\partial u^j}, \quad (19.61)$$

where the second term arises because, in general, the basis vectors  $\mathbf{e}_i$  are not constant. Using (19.36), we write

$$\frac{\partial \mathbf{v}}{\partial u^j} = \frac{\partial v^i}{\partial u^j} \mathbf{e}_i + v^i \Gamma_{ij}^k \mathbf{e}_k.$$

Since  $i$  and  $k$  are dummy indices, we may interchange them to obtain

$$\begin{aligned} \frac{\partial \mathbf{v}}{\partial u^j} &= \frac{\partial v^i}{\partial u^j} \mathbf{e}_i + v^k \Gamma_{kj}^i \mathbf{e}_i \\ &= \left( \frac{\partial v^i}{\partial u^j} + v^k \Gamma_{kj}^i \right) \mathbf{e}_i. \end{aligned} \quad (19.62)$$

The quantity in parentheses is referred to specifically as the **covariant derivative of a vector**:

♠ **Covariant derivative of a vector:**

The quantities defined by

$$v^i_{;j} \equiv \frac{\partial v^i}{\partial u^j} + \Gamma_{kj}^i v^k. \quad (19.63)$$

are called **covariant derivatives** of contravariant components  $v^i$  of a vector  $\mathbf{v}$  with respect to  $u^j$ . Here, the semicolon subscript on the left-hand side denotes covariant differentiation.

Using this notation, we may write the derivative of a vector in the very compact form

$$\frac{\partial \mathbf{v}}{\partial u^j} = v^i_{;j} \mathbf{e}_i.$$

The corresponding result for the *covariant* components  $v_i$  can be found in a similar way by considering the derivative of  $\mathbf{v} = v_i \mathbf{e}^i$  and using (19.38) to obtain

$$v_i_{;j} = \frac{\partial v_i}{\partial u^j} - \Gamma_{ij}^k v_k. \quad (19.64)$$

#### 19.4.2 Remarks on Covariant Derivatives

1. The arrangement of indices  $i, j, k$  in the Christoffel symbols in (19.63) and (19.64) can be determined systematically in the following manner. First, the index to which the derivative is taken (i.e.,  $j$  in this case) is the last subscript on the Christoffel symbol. Secondly, the other index appearing on the left-hand side (i.e.,  $i$  in this case) also appears in the Christoffel symbol on the right-hand side without raising or lowering. The remaining index can then be arranged in only one.

2. Similar to  $v^i_{;j}$ , a comparable short-hand notation for partial derivatives is obtained by replacing the semicolon by a comma such as

$$v^i_{,j} \equiv \frac{\partial v^i}{\partial u^j} \quad \text{and} \quad v_{i,j} \equiv \frac{\partial v_i}{\partial u^j}.$$

3. In Cartesian coordinates, all the  $\Gamma^i_{kj}$  are zero, so the covariant derivative reduces to the simple partial derivative, say,  $v^i_{;j} \equiv v^i_{,j}$ .

### 19.4.3 Covariant Derivatives of Tensors

Covariant derivatives of higher-order tensors can be defined by a procedure similar to the one for vectors. As an example, let us consider the derivative of the second-order tensor  $\mathbf{T}$  with respect to the coordinate  $u^k$ . Expressing  $\mathbf{T}$  in terms of its contravariant components, we have

$$\begin{aligned} \frac{\partial \mathbf{T}}{\partial u^k} &= \frac{\partial}{\partial u^k} (T^{ij} \mathbf{e}_i \otimes \mathbf{e}_j) \\ &= \frac{\partial T^{ij}}{\partial u^k} \mathbf{e}_i \otimes \mathbf{e}_j + T^{ij} \frac{\partial \mathbf{e}_i}{\partial u^k} \otimes \mathbf{e}_j + T^{ij} \mathbf{e}_i \otimes \frac{\partial \mathbf{e}_j}{\partial u^k}. \end{aligned} \quad (19.65)$$

Using Christoffel symbols, we obtain

$$\frac{\partial \mathbf{T}}{\partial u^k} = \frac{\partial T^{ij}}{\partial u^k} \mathbf{e}_i \otimes \mathbf{e}_j + T^{ij} \Gamma_{ik}^\ell \mathbf{e}_\ell \otimes \mathbf{e}_j + T^{ij} \mathbf{e}_i \otimes \Gamma_{jk}^\ell \mathbf{e}_\ell.$$

Interchanging the dummy indices  $i$  and  $l$  in the second term and  $j$  and  $l$  in the third term on the right-hand side, we set

$$\frac{\partial \mathbf{T}}{\partial u^k} = \left( \frac{\partial T^{ij}}{\partial u^k} + \Gamma_{lk}^i T^{lj} + \Gamma_{lk}^j T^{il} \right) \mathbf{e}_i \otimes \mathbf{e}_j,$$

where the expression in parentheses is the required covariant derivative defined by

$$T^{ij}_{;k} = \frac{\partial T^{ij}}{\partial u^k} + \Gamma_{lk}^i T^{lj} + \Gamma_{lk}^j T^{il}. \quad (19.66)$$

Using the notation (19.66), we can write the derivative of the tensor  $\mathbf{T}$  with respect to  $u^k$  as

$$\frac{\partial \mathbf{T}}{\partial u^k} = T^{ij}_{;k} \mathbf{e}_i \otimes \mathbf{e}_j.$$

Results similar to (19.66) can be obtained for the covariant derivatives of the mixed and covariant components of a second-order tensor. Collecting all of these results leads to the following:

♠ **Covariant derivative of a tensor:**

Covariant derivatives of components of a second-order tensor  $\mathbf{T}$  are given by

$$\begin{aligned} T^{ij}_{;k} &= T^{ij}_{,k} + \Gamma_{lk}^i T^{lj} + \Gamma_{lk}^j T^{il}, \\ T^i_{j;k} &= T^i_{j,k} + \Gamma_{lk}^i T_j^l - \Gamma_{jk}^\ell T_\ell^i, \\ T_{ij;k} &= T_{ij,k} - \Gamma_{ik}^\ell T_{lj} - \Gamma_{jk}^\ell T_{il}, \end{aligned}$$

where the comma notation means the taking of partial derivatives.

The position of the indices in the expressions is very systematic. We focus on the index  $i$  or  $j$  on the left-hand side. First, the index  $k$  to which the derivative is taken should be the last subscript on the Christoffel symbol. Next, if the index ( $i$  or  $j$ ) on the left-hand side is a superscript, then the corresponding term on the right-hand side containing a Christoffel symbol is attached to a plus sign. In contrast, when the index on the left-hand side is a subscript, the corresponding term on the right is attached to a minus sign. We can extend this in a straightforward manner to tensors with an arbitrary number of contravariant and covariant indices.

*Remark.*

1. All of the quantities  $T^{ij}_{;k}$ ,  $T^i_{j;k}$ , and  $T_{ij;k}$  are the components of the same third-order tensor  $\nabla \mathbf{T}$  with respect to different tensor bases, i.e.,

$$\nabla \mathbf{T} = T^{ij}_{;k} \mathbf{e}_i \otimes \mathbf{e}_j \otimes \mathbf{e}^k = T^i_{j;k} \mathbf{e}_i \otimes \mathbf{e}^j \otimes \mathbf{e}^k = T_{ij;k} \mathbf{e}^i \otimes \mathbf{e}^j \otimes \mathbf{e}^k.$$

2. In general, we may call the  $v^i_{;j}$  the covariant derivative of  $\mathbf{v}$  and denote it by  $\nabla \mathbf{v}$ . In Cartesian coordinates, its components are just  $\partial v^i / \partial x^j$ .
3. Given a metric tensor  $\mathbf{g}$ , the covariant derivatives of its components,  $g_{ij;k}$  and  $g^{ij}_{;k}$ , are identically zero in terms of arbitrary coordinates. This is called **Ricci's theorem**, for which we give the proof in Exercises 2 and 3).

#### 19.4.4 Vector Operators in Tensor Form

This subsection is devoted to finding expressions for vector differential operators such as grad, div, rot, and the Laplacian in tensor form that are valid in general coordinate systems. In principle, they are obtained in a straightforward manner by replacing the partial derivative given in Cartesian coordinates with covariant derivatives. These tensor forms, however, can be simplified by using the metric tensor  $g_{ij}$  as shown below.

- 1. Gradient:** The gradient of a scalar  $\phi$  in a general coordinate system is given by

$$\nabla\phi = \phi_{;i} \mathbf{e}^i = \frac{\partial\phi}{\partial u^i} \mathbf{e}^i, \quad (19.67)$$

since the covariant derivative of a scalar is the same as its partial derivative.

- 2. Divergence:** The tensor form of the divergence of a vector  $\mathbf{v}$  is given by

$$\nabla \cdot \mathbf{v} = v^i_{;i} = \frac{\partial v^i}{\partial u^i} + \Gamma^i_{ki} v^k. \quad (19.68)$$

Observe that the index  $i$  appears twice in the Christoffel symbol. Using the expression (see Sect. 19.3.3)

$$\Gamma^k_{ik} = \frac{\partial}{\partial u^i} \log \sqrt{g} = \frac{1}{\sqrt{g}} \frac{\partial \sqrt{g}}{\partial u^i},$$

we obtain a more compact form:

$$\begin{aligned} v^i_{;i} &= \frac{\partial v^i}{\partial u^i} + \Gamma^i_{ki} v^k = \frac{\partial v^i}{\partial u^i} + \left( \frac{1}{\sqrt{g}} \frac{\partial \sqrt{g}}{\partial u^k} \right) v^k \\ &= \frac{1}{\sqrt{g}} \sqrt{g} \left( \frac{\partial v^i}{\partial u^i} \right) + \frac{1}{\sqrt{g}} \left( \frac{\partial \sqrt{g}}{\partial u^k} \right) v^k = \frac{1}{\sqrt{g}} \frac{\partial}{\partial u^k} (\sqrt{g} v^k). \end{aligned} \quad (19.69)$$

- 3. Laplacian:** The tensor form of the Laplacian  $\nabla^2\phi$  is obtained by making use of the following relation:

$$v^i_{;i} = \nabla \cdot \mathbf{v} = \nabla \cdot (\nabla\phi) = \nabla^2\phi,$$

where we assume that  $\mathbf{v} = \nabla\phi$ . From (19.67), we have

$$v_i \mathbf{e}^i = \mathbf{v} = \nabla\phi = \frac{\partial\phi}{\partial u^i} \mathbf{e}^i.$$

Thus the covariant components of  $\mathbf{v}$  are given by

$$v_i = \frac{\partial\phi}{\partial u^i},$$

and its contravariant components  $v^i$  can be obtained by raising the index using the metric tensor:

$$v^j = g^{jk} v_k = g^{jk} \frac{\partial\phi}{\partial u^k}.$$

Substituting this into (19.69), we finally arrive at

$$\nabla^2\phi = v^j_{;j} = \frac{1}{\sqrt{g}} \frac{\partial}{\partial u^j} \left( \sqrt{g} g^{jk} \frac{\partial\phi}{\partial u^k} \right). \quad (19.70)$$

4. **Rotation:** In general curvilinear coordinates, the operation  $\nabla \times \mathbf{v}$  is defined by

$$[\nabla \times \mathbf{v}]_{ij} = v_{i;j} - v_{j;i}, \quad (19.71)$$

which forms covariant components of an antisymmetric tensor. The right-hand side of (19.71) can be simplified as

$$\begin{aligned} v_{i;j} - v_{j;i} &= \frac{\partial v_i}{\partial u^j} - \Gamma_{ij}^\ell v_\ell - \frac{\partial v_j}{\partial u^i} + \Gamma_{ji}^\ell v_\ell \\ &= \frac{\partial v_i}{\partial u^j} - \frac{\partial v_j}{\partial u^i}, \end{aligned} \quad (19.72)$$

where the Christoffel symbols cancel out owing to their symmetric properties. Therefore, components of the tensor  $\nabla \times \mathbf{v}$  can be written in terms of partial derivatives as

$$[\nabla \times \mathbf{v}]_{ij} = \frac{\partial v_i}{\partial u^j} - \frac{\partial v_j}{\partial u^i}. \quad (19.73)$$

Our results are summarized as follows:

♠ **Vector operators in tensor forms:**

$$\begin{aligned} 1. \quad \nabla \phi &= \frac{\partial \phi}{\partial u^i} \mathbf{e}^i. & 2. \quad \nabla \cdot \mathbf{v} &= \frac{1}{\sqrt{g}} \frac{\partial}{\partial u^i} (\sqrt{g} v^i). \\ 3. \quad \nabla^2 \phi &= \frac{1}{\sqrt{g}} \frac{\partial}{\partial u^i} \left( \sqrt{g} g^{ik} \frac{\partial \phi}{\partial u^k} \right). & 4. \quad [\nabla \times \mathbf{v}]_{ij} &= \frac{\partial v_i}{\partial u^j} - \frac{\partial v_j}{\partial u^i}. \end{aligned}$$

## Exercises

1. Prove that the covariant derivatives  $v^k_{;j}$  form a second-order tensor of type  $(1, 1)$ .

**Solution:** Employ the transformation laws of  $v^k$  and  $\Gamma_{pj}^k$  [see (19.44)] to obtain

$$\begin{aligned} v^k_{;j} &= \frac{\partial v^k}{\partial u^j} + \Gamma_{pj}^k v^p \\ &= \frac{\partial}{\partial u^j} \left( \frac{\partial u^k}{\partial u'^q} v'^q \right) + \left( \frac{\partial u^k}{\partial u'^q} \frac{\partial u'^r}{\partial u^p} \frac{\partial u'^s}{\partial u^j} \Gamma'^q_{rs} + \frac{\partial u^k}{\partial u'^q} \frac{\partial^2 u'^q}{\partial u^p \partial u^j} \right) \left( \frac{\partial u^p}{\partial u'^t} v'^t \right) \\ &= \frac{\partial^2 u^k}{\partial u'^a \partial u'^q} \frac{\partial u'^a}{\partial u^j} v'^q + \frac{\partial u^k}{\partial u'^q} \frac{\partial u'^t}{\partial u^j} \frac{\partial v'^q}{\partial u'^t} \\ &\quad + \frac{\partial u^k}{\partial u'^q} \left( \frac{\partial u'^r}{\partial u^p} \frac{\partial u'^s}{\partial u^j} \Gamma'^q_{rs} + \frac{\partial^2 u'^q}{\partial u^p \partial u^j} \right) \frac{\partial u^p}{\partial u'^t} v'^t. \end{aligned} \quad (19.74)$$

The sum of the terms involving second derivatives is zero; this is seen by taking a partial derivative with respect to  $u'^q$  in the expression

$$\frac{\partial u^k}{\partial u'^a} \frac{\partial u'^a}{\partial u^j} = \delta_j^k,$$

which yields

$$\begin{aligned} \frac{\partial}{\partial u'^q} \left( \frac{\partial u^k}{\partial u'^a} \frac{\partial u'^a}{\partial u^j} \right) &= \frac{\partial^2 u^k}{\partial u'^q \partial u'^a} \frac{\partial u'^a}{\partial u^j} + \frac{\partial u^k}{\partial u'^a} \frac{\partial^2 u'^a}{\partial u'^q \partial u^j} \\ &= \frac{\partial^2 u^k}{\partial u'^q \partial u'^a} \frac{\partial u'^a}{\partial u^j} + \frac{\partial u^k}{\partial u'^a} \frac{\partial u^m}{\partial u'^q} \frac{\partial^2 u'^a}{\partial u^m \partial u^j} \\ &= 0. \end{aligned} \quad (19.75)$$

From (19.74) and (19.75), it follows that

$$\begin{aligned} v^k_{;j} &= \frac{\partial u^k}{\partial u'^q} \frac{\partial u'^t}{\partial u^j} \frac{\partial v'^q}{\partial u'^t} + \frac{\partial u^k}{\partial u'^q} \delta'^r_t \frac{\partial u'^s}{\partial u^j} \Gamma'^q_{rs} v'^t \\ &= \frac{\partial u^k}{\partial u'^q} \frac{\partial u'^s}{\partial u^j} \left( \frac{\partial v'^q}{\partial u'^s} + \Gamma'^q_{ts} v'^t \right) \\ &= \frac{\partial u^k}{\partial u'^q} \frac{\partial u'^s}{\partial u^j} v'^q_{;s}, \end{aligned}$$

in which the last term in the last line,  $v'^q_{;s}$ , represents the covariant derivative of  $v'^q$  with respect to the *primed* coordinates  $u'^s$ . Hence, we see that the  $v^k_{;j}$  form a second-order tensor of type (1, 1). ♣

2. Show that the metric tensor is a **covariant constant**, i.e., the covariant derivative of any component is identically zero:  $g_{kp};_j = 0$ . This result is known as **Ricci's theorem**.

**Solution:** It follows that

$$\begin{aligned} g_{kp};_j &= g_{kp,j} - \Gamma^r_{kj} g_{rp} - \Gamma^r_{pj} g_{kr} \\ &= g_{kp,j} - \frac{1}{2} \delta_p^s (g_{js,k} + g_{sk,j} - g_{kj,s}) - \frac{1}{2} \delta_k^s (g_{js,p} + g_{sp,j} - g_{pj,s}) \\ &= 0. \quad \clubsuit \end{aligned}$$

3. Show that  $\delta_k^p$  and  $g^{kp}$  are also covariant constants.

**Solution:** We have  $\delta_k^p_{;j} = \delta_{k,j}^p - \Gamma^p_{qj} \delta_k^q - \Gamma^q_{kj} \delta_q^p = 0$ , which completes our first proof. Next, observe that  $\delta_j^k = g_{jp} g^{pk}$  to find the identity

$$0 = \delta_j^k_{;q} = (g_{jp} g^{pk})_{;q} = g_{jp;q} g^{pk} + g_{jp} g^{pk}_{;q}.$$

Since  $g_{jp}$  is a covariant constant, the first term in the last expression has the value zero. Multiplication by  $g^{jr}$  produces the desired result. ♣

*Remark.* Owing to Ricci's theorem and its two corollaries noted above, the components of the metric tensor can be regarded as constants under covariant differentiation. Thus, e.g.,

$$\begin{aligned} g_{il} A^\ell{}_{;k} &= (g_{il} A^\ell)_{;k} = A_{i;k}, \\ g_{il} T^{\ell m}{}_{;k} &= (g_{il} T^{\ell m})_{;k} = T_i{}^m{}_{;k}, \\ T_{ik;\ell} g^{im} g^{kn} &= (T_{ik} g^{im} g^{kn})_{;\ell} = T^{mn}{}_{;\ell}, \end{aligned}$$

and so on.

3. Use (19.70) to find expressions for  $\nabla^2 \phi$  and  $\nabla \cdot \mathbf{v}$  in an orthogonal coordinate system with scale factors  $h_i$  ( $i = 1, 2, 3$ ).

**Solution:** For an orthogonal coordinate system  $\sqrt{g} = h_1 h_2 h_3$ ; further,  $g^{ii} = 1/h_i^2$  for fixed  $i$  and  $g^{ij} = 0$  for  $i \neq j$ . Therefore, from (19.70), we set

$$\nabla^2 \phi = \frac{1}{h_1 h_2 h_3} \frac{\partial}{\partial u^j} \left( \frac{h_1 h_2 h_3}{h_j^2} \frac{\partial \phi}{\partial u^j} \right).$$

In a similar manner, we have

$$\nabla \cdot \mathbf{v} = \frac{1}{h_1 h_2 h_3} \frac{\partial}{\partial u^j} (h_1 h_2 h_3 v^j). \quad \clubsuit$$

## 19.5 Applications in Physics and Engineering

### 19.5.1 General Relativity Theory

It cannot be denied that the **general relativity theory** is one of the most famous and beautiful applications of non-Cartesian tensor calculus in physics. This section outlines the concepts one needs in order to understand the general theory of relativity, which is necessary for obtaining the gravitational field equation and relevant tensorial quantities that are involved with the equation.

Before proceeding to the argument, let us point out that the notion of **geometric curvature** is central to general relativity, which quantifies the curvature of space at any given point in the space considered. In Sect. 19.2.3, we learned that a space is a **flat** locally (or entirety), if there exist coordinates  $x^i$  such that the line element through a limited region (or the whole) can be written as

$$(ds)^2 = \varepsilon_i (dx^i)^2,$$

where  $\varepsilon = \pm 1$ . However, if we employ a different coordinate system  $x'^i$ , the line element  $(ds)^2$ , in general, is not of the above form, but reads as

$$(ds)^2 = g_{ij} dx^i dx^j$$

with the appropriate metric tensor  $g_{ij}$ . Hence, we require a means of identifying a flat space directly from the metric  $g_{ij}$ , *independent* of our choice of coordinate system. Such a coordinate-independent way of defining the curvature of a space leads to the field equation of gravity, i.e., **Einstein's field equation**, described in Sect. 19.5.4.

### 19.5.2 Riemann Tensor

The curvature of space can be quantified in a manner independent of the coordinate system by changing the order of **covariant differentiation**. Covariant differentiation is a generalization of partial differentiation, in which interchanging the order of differentiation changes the result. To illustrate this, let us consider an arbitrary vector field with covariant components  $v_i$ . The covariant derivative of  $v_i$  is given by [see (19.64)]

$$v_{i;j} = \frac{\partial v_i}{\partial u^j} - \Gamma_{ij}^\ell v_\ell.$$

A second covariant differentiation then yields

$$\begin{aligned} (v_{i;j})_{;k} &= \frac{\partial v_{i;j}}{\partial u^k} - \Gamma_{ik}^m v_{m;j} - \Gamma_{jk}^m v_{i;m} \\ &= \frac{\partial^2 v_i}{\partial u^j \partial u^k} - \left( \frac{\partial \Gamma_{ij}^\ell}{\partial u^k} \right) v_\ell - \Gamma_{ij}^\ell \left( \frac{\partial v_\ell}{\partial u^k} \right) \\ &\quad - \Gamma_{ik}^m \left( \frac{\partial v_m}{\partial u^j} - \Gamma_{mj}^\ell v_\ell \right) - \Gamma_{jk}^m \left( \frac{\partial v_i}{\partial u^m} - \Gamma_{im}^\ell v_\ell \right). \end{aligned}$$

By interchanging the indices  $j$  and  $k$  to obtain the expression corresponding to  $(v_{i;k})_{;j}$  and then subtracting the expression we set from the above relation gives us

$$(v_{i;j})_{;k} - (v_{i;k})_{;j} = R_{ijk}^\ell v_\ell,$$

where

$$R_{ijk}^\ell = \frac{\partial \Gamma_{ik}^\ell}{\partial u^j} - \frac{\partial \Gamma_{ij}^\ell}{\partial u^k} + \Gamma_{ik}^m \Gamma_{mj}^\ell - \Gamma_{ij}^m \Gamma_{mk}^\ell. \quad (19.76)$$

The quantity  $R_{ijk}^\ell$  shown on the left-hand side is called the **Riemann tensor** (or **curvature tensor**). Since Christoffel symbols  $\Gamma_{ij}^k$  are functions of the metric tensor  $g_{ij}$ , (19.76) indicates that the Riemann tensor is defined in terms of the metric tensor and its first and second derivatives.

Recall that if the space being considered is flat, we may choose coordinates such that  $\Gamma_{ij}^k$  and its derivatives vanish. Therefore, we have

$$R_{ijk}^\ell = 0 \quad (19.77)$$

at every point in the flat region. In fact, it is possible to show that (19.77) is a necessary and sufficient condition for the region of a space to be flat. Consequently, we conclude the following: when the Riemann tensor satisfies (19.77), it indicates that the region of a space is flat and when it does not satisfy (19.77), the region is curved.

Two relevant quantities are obtained by contracting the Riemann tensor. One is the **Ricci tensor** defined by

$$R_{ij} \equiv R_{ijk}^k$$

and the other is the **scalar curvature** (or **Ricci scalar**) given by

$$R \equiv g^{ij} R_{ij} = R_i^i.$$

These two quantities are important for introducing the **Einstein tensor**

$$G^{ij} = R^{ij} - \frac{1}{2}g^{ij}R,$$

which describes the space-time curvature in the field equation of general relativity.

### 19.5.3 Energy–Momentum Tensor

We now wish to determine the form of the gravitational field equation that, in the weak limit of a static gravitational field, reduces, to the classical **Newtonian field of gravity** described by

$$\nabla^2\Phi = 4\pi G\rho. \quad (19.78)$$

Here,  $\Phi$  is the **potential field** that corresponds to the space-time curvature in relativistic theory,  $G$  is the **universal gravitational constant**, and  $\rho$  is the mass-density distribution of matter. Note that (19.78) is a form of **Poisson's equation** with  $4\pi G\rho$  as the source term. This implies the presence of a corresponding source term associated with the space-time curvature in Einstein's field equation. This source term is given by the **energy-momentum tensor**  $T^{ij}$  defined by

$$T^{ij} = \rho u^i u^j.$$

Here,  $\rho$  is the density of matter,  $u^i$  is the **four-velocity** represented by  $u^i = (u^0, u^1, u^2, u^3) = (\gamma c, \gamma \mathbf{v})$ , where  $c$  is the velocity of light,  $\mathbf{v}$  is the three-dimensional velocity (nonrelativistic) of a particle, and  $\gamma = (1 - v^2/c^2)^{-1/2}$ .

The physical interpretations of the components of the energy-momentum tensor are:

$T^{00}$  : the energy density of the particles.

$T^{0i}$  : the energy flux (the heat conduction) in the  $i$ th direction.

$T^{i0}$  : the momentum density in the  $i$ th direction.

$T^{ij}$  : the flow of the  $i$ th-component momentum in the  $j$ th direction  
(i.e., the random thermal motions giving rise to viscous stress).

#### 19.5.4 Einstein Field Equation

The parameters necessary to obtain Einstein's field equation, which relates the geometric space-time curvature to the density of mass-energy, are already on hand. One side of the equation should comprise the measure of the density of mass-energy, i.e., the stress-energy tensor  $T_{ij}$ , and the other side should consist of a measure of the curvature involving the Ricci curvature  $R_{ij}$  and scalar curvature  $R$ . By making this equation consistent with Newton's equation of motion in the limit of a weak gravitational force as well as with several postulates from a physical standpoint, Einstein's field equation is obtained in the following form:

$$R_{ij} - \frac{1}{2}g_{ij}R = \frac{8\pi G}{c^4}T_{ij}. \quad (19.79)$$

Given the matter source  $T_{ij}$ , this tensor equation is composed of ten partial differential equations for the metric tensor  $g_{ij}(x)$ . Apparently, the tensor equation is analogous to the **Maxwell equations** that determine the electromagnetic field given the charge and current densities (see Sect. 18.5.3). Unlike the Maxwell equations, however, the differential equations of gravitational theory are nonlinear, which make them very difficult to solve. Surprisingly, despite the nonlinearity, a number of exact solutions have been obtained owing to the presence of symmetries in space-time, which restrict the possible forms of the metric.

**Remark.** Einstein's field equation in (19.79) is the most fundamental equation in classical physics. The explicit form of the equation can be derived from a few arguments. However, it cannot be derived from other physical principles since there is no theory that is more fundamental.



## Tensor as Mapping

**Abstract** In this chapter, we show that tensors can be identified with mathematical operators that transform elements from one abstract vector space to another. This viewpoint on tensors is apparently different from those presented in Chaps. 18 and 19, where tensors have been identified as sets of index quantities subject to a transformation law under changes of coordinate systems. However, the viewpoint presented here turns out to be consistent with those presented in the previous two chapters when we introduce the concept of inner product into the abstract vector space (Sect. 20.3.4).

### 20.1 Vector as a Linear Function

#### 20.1.1 Overview

In Chaps. 19 and 20 tensors are defined as collections of index quantities that obey characteristic transformation laws under a change of coordinate systems. In this chapter we present an alternative definition of tensors; that does not require specification of a coordinate system, so that it is suitable for more general tensor analyses describing geometric properties of abstract vector spaces other than our familiar three-dimensional Euclidean space.

The crucial point is that in this alternative definition, a tensor is considered not as a set of index quantities but as an operator (linear function or mapping) acting on vector spaces. For instance, a second-order tensor  $\mathbf{T}$  is identified with a linear function that associates two vectors  $\mathbf{v}$  and  $\mathbf{w}$  with a real number  $c \in \mathbf{R}$ , which is symbolized by

$$\mathbf{T}(\mathbf{v}, \mathbf{w}) = c.$$

Emphasis should be placed on the fact that such a generalized definition of tensors applies to all kinds of general vector spaces (finite-dimensional), regardless of whether or not they possess geometric properties such as the **distance**, **norm**, or **inner product** of their elements (see Sect. 4.2.1). In

fact, the tensors we discussed earlier belong to a specific class of more general tensors, in the sense that they were defined solely on the three-dimensional Euclidean space, a particular class of vector spaces endowed with the inner product property. However, we shall see that, the concept of tensor can be extended beyond inner product spaces by introducing the more general definition referred to above.

Throughout the following discussion, we restrict our arguments to finite-dimensional vector spaces over  $\mathbf{R}$  in order to provide a minimal course for general tensor calculus.

### 20.1.2 Vector Spaces Revisited

To begin with, we briefly review the definition of abstract vector spaces. A **vector space** (or **linear space**)  $V$  over  $\mathbf{R}$  is a set of elements called **vectors** that have two operations, **addition** and **scalar multiplication**, and a distinguishing element  $\mathbf{0} \in V$ . Here, addition (denoted by  $+$ ) assigns to each pair of elements  $\mathbf{v}, \mathbf{w} \in V$  a third element  $\mathbf{v} + \mathbf{w} \in V$  and the scalar multiplication assigns an element  $c\mathbf{v} \in V$  to each  $\mathbf{v} \in V$  and  $c \in \mathbf{R}$ . By definition, all of the elements  $\mathbf{v}, \mathbf{w}, \mathbf{x} \in V$  and all  $a, b \in \mathbf{R}$  must satisfy the following axioms.

1. The commutative law for  $+$ , i.e.,  $\mathbf{v} + \mathbf{w} = \mathbf{w} + \mathbf{v}$ .
2. The associative law for  $+$ , i.e.,  $(\mathbf{v} + \mathbf{w}) + \mathbf{x} = \mathbf{v} + (\mathbf{w} + \mathbf{x})$ .
3. Existence of identity for  $+$ , i.e.,  $\mathbf{v} + \mathbf{0} = \mathbf{v}$ .
4. Existence of negatives, i.e., there is  $-\mathbf{v}$  such that  $\mathbf{v} + (-\mathbf{v}) = \mathbf{0}$ .
5.  $a(\mathbf{v} + \mathbf{w}) = a\mathbf{v} + a\mathbf{w}$ .
6.  $(a + b)\mathbf{v} = a\mathbf{v} + b\mathbf{v}$ .
7.  $(ab)\mathbf{v} = a(b\mathbf{v})$ .
8.  $1\mathbf{v} = \mathbf{v}$ .

Given two vector spaces  $V$  and  $W$ , it is possible to set a function  $f$  so that

$$f: V \rightarrow W.$$

The function  $f$  is called a **linear function** (or **linear mapping**) of  $V$  into  $W$  if for all  $\mathbf{v}_1, \mathbf{v}_2 \in V$  and  $c \in \mathbf{R}$  it yields

$$\begin{aligned} f(\mathbf{v}_1 + \mathbf{v}_2) &= f(\mathbf{v}_1) + f(\mathbf{v}_2), \\ f(c\mathbf{v}_1) &= cf(\mathbf{v}_1). \end{aligned}$$

### 20.1.3 Vector Spaces of Linear Functions

In elementary calculus, the concepts of vectors and linear functions are distinguished from one another: vectors are elements of a vector space and linear functions provide a correspondence between them. However, in view of the

axioms **1** to **8** above, we observe that the set of linear functions  $f, g, \dots$  of  $V$  into  $W$  also forms a vector space in which addition and scalar multiplication, respectively, are defined by

$$(f + g)\mathbf{v} = f(\mathbf{v}) + g(\mathbf{v}), \quad (20.1)$$

and

$$(cf)\mathbf{v} = cf(\mathbf{v}), \quad (20.2)$$

where  $\mathbf{v} \in V$  and  $f(\mathbf{v}), g(\mathbf{v}) \in W$ . We denote by  $\mathcal{L}(V, W)$  a vector space spanned by a set of linear functions  $f$  as

$$f : V \rightarrow W.$$

It is a trivial matter to verify that  $f + g$  and  $cf$  are also linear functions and so belong to the same vector space  $\mathcal{L}(V, W)$ . These arguments are summarized by the following important theorem:

♠ **Vector space of linear functions:**

Let  $V$  and  $W$  be vector spaces. A set of linear functions  $f : V \rightarrow W$  forms a vector space denoted by  $\mathcal{L}(V, W)$ .

This theorem states that the linear functions  $f_1, f_2, \dots$  of  $V$  into  $W$  are elements of a vector space  $\mathcal{L}(V, W)$ , analogous to vectors  $\mathbf{v}_1, \mathbf{v}_2, \dots$  being elements of a vector space  $V$ . This analogy implies that a linear function  $f \in \mathcal{L}(V, W)$  can be regarded as a *vector* and, conversely, that a vector  $\mathbf{v} \in V$  can be regarded as a *linear function*. Such identifying vectors and linear functions is crucially important for obtaining a generalized definition of tensors that is free of the concept of inner product and the specification of a coordinate system.

#### 20.1.4 Dual Spaces

Let  $V^*$  denote a set of all linear functions such as

$$f : V \rightarrow \mathbf{R}.$$

(Note that the asterisk (\*) in  $V^*$  does not mean complex conjugate.) Then, since

$$V^* = \mathcal{L}(V, \mathbf{R}),$$

it follows that  $V^*$  is a vector space. The vector space  $V^*$  is called the **dual space** (or **conjugate space**) of  $V$ , whose elements  $f \in V^*$  associate a vector  $\mathbf{v} \in V$  with a real number  $c \in \mathbf{R}$ , symbolized as

$$f(\mathbf{v}) = c.$$

Particularly important elements of  $V^*$  are linear functions

$$\varepsilon^i : V \rightarrow \mathbf{R} \quad (i = 1, 2, \dots, n)$$

that associate a **basis** vector  $\mathbf{e}_i \in V$  with the unit number 1. In fact, a set of such linear functions  $\{\varepsilon^j\}$  serves as a basis of the dual space  $V^*$  as stated below.

♠ **Dual basis:**

For each basis  $\{\mathbf{e}_i\}$  for  $V$ , there is a unique basis  $\{\varepsilon^j\}$  for  $V^*$  such that

$$\varepsilon^j(\mathbf{e}_i) = \delta_i^j. \quad (20.3)$$

The linear functions  $\varepsilon^j : V \rightarrow \mathbf{R}$  defined by (20.3) make up the **dual basis** to the basis  $\{\mathbf{e}_i\}$  of  $V$ .

*Proof* Let us verify that the set of  $\{\varepsilon^j\}$  defined by (20.3) serves as a basis of  $V^*$ . Recall that in finite dimensions, a basis of a vector space  $V$  is defined as a set of linearly independent vectors that spans all of  $V$ . To show linear independence, we assume that  $a_j \varepsilon^j = 0$ . Then we have

$$a_j \varepsilon^j(\mathbf{e}_i) = a_j \delta_i^j = a_i = 0 \quad \text{for all } i,$$

which implies that  $\{\varepsilon^j\}$  is linearly independent. ♣

**Remark.** Raising of the index  $j$  attached to  $\varepsilon^j$  is intentional, as this convention is necessary to provide a consistent notation of components of generalized tensors, demonstrated in Sect. 20.3.

*Examples* Expand a vector  $\mathbf{v} \in V$  as

$$\mathbf{v} = v^i \mathbf{e}_i,$$

to find that

$$\varepsilon^j(\mathbf{v}) = \varepsilon^j(v^i \mathbf{e}_i) = v^i \varepsilon^j(\mathbf{e}_i) = v^i \delta_i^j = v^j.$$

This indicates that  $\varepsilon^j$  is the linear function that scans the  $j$ th component of  $\mathbf{v}$  with respect to the basis  $\{\mathbf{e}_i\}$ .

### 20.1.5 Equivalence Between Vectors and Linear Functions

If  $V$  is a vector space and  $\tau \in V^*$ , then  $\tau$  is a function of the variable  $\mathbf{v} \in V$  that generates a real number denoted by  $\tau(\mathbf{v})$ . Owing to the identification of vectors and linear functions, however, it is possible to reverse our reasoning and consider  $\mathbf{v}$  as a *function* of the *variable*  $\tau$ , again with the real value  $\mathbf{v}(\tau) = \tau(\mathbf{v})$ . When we take this approach,  $\mathbf{v}$  is a linear function on  $V^*$ .

*Remark.* The two views contrasted above are both asymmetric, but this asymmetry can be eliminated by introducing the notation

$$\langle \ , \ \rangle : V \times V^* \rightarrow \mathbf{R},$$

which gives

$$\langle \mathbf{v}, \tau \rangle = \tau(\mathbf{v}) = \mathbf{v}(\tau) \in \mathbf{R}.$$

Here  $\langle \ , \ \rangle$  is a function of two variables  $\mathbf{v}$  and  $\tau$ , called the **natural pairing** of  $V$  and  $V^*$  into  $\mathbf{R}$ . It is easy to verify that  $\langle \ , \ \rangle$  is bilinear.

The concepts and notation introduced in Sect. 20.1.3 and 20.1.4 serve as preliminaries for the discussions in the following sections.

## 20.2 Tensor as Multilinear Function

### 20.2.1 Direct Product of Vector Spaces

To arrive at the new definition of tensors we are seeking requires three more concepts, demonstrated in Sect. 20.2.1–20.2.2.

The first is the direct product of vector spaces; if  $V$  and  $W$  are vector spaces, then we can establish a new vector space by forming the **direct product** (or **Cartesian product**)  $V \times W$  of the two spaces. The direct product  $V \times W$  consists of *ordered* pairs  $(\mathbf{v}, \mathbf{w})$  with  $\mathbf{v} \in V$  and  $\mathbf{w} \in W$ , as symbolized by

$$V \times W = \{(\mathbf{v}, \mathbf{w}) \mid \mathbf{v} \in V, \mathbf{w} \in W\}.$$

The addition and scalar multiplication of the elements are defined by

$$\begin{aligned} (\mathbf{v}, \mathbf{w}_1) + (\mathbf{v}, \mathbf{w}_2) &= (\mathbf{v}, \mathbf{w}_1 + \mathbf{w}_2), \\ (\mathbf{v}_1, \mathbf{w}) + (\mathbf{v}_2, \mathbf{w}) &= (\mathbf{v}_1 + \mathbf{v}_2, \mathbf{w}), \\ c(\mathbf{v}, \mathbf{w}) &= (c\mathbf{v}, \mathbf{w}) = (\mathbf{v}, c\mathbf{w}). \end{aligned}$$

The linear dimension of the resulting vector spaces  $V \times W$  equals the product of the linear dimensions of  $V$  and  $W$ . The elements  $(\mathbf{v}, \mathbf{w})$  of the direct product  $V \times W$  is sometimes noted by  $\mathbf{vw}$ .

*Remark.* The reader should note a distinction between the direct product  $V \times W$  and the **direct sum**  $V + W$  of the two vector spaces. A direct sum  $V + W$  consists of all pairs  $(\mathbf{v}, \mathbf{w}) \equiv (\mathbf{w}, \mathbf{v})$  with  $\mathbf{v} \in V$  and  $\mathbf{w} \in W$  for which addition and scalar multiplication are defined by

$$(\mathbf{v}_1, \mathbf{w}_1) + (\mathbf{v}_2, \mathbf{w}_2) = (\mathbf{v}_1 + \mathbf{v}_2, \mathbf{w}_1 + \mathbf{w}_2), \quad c(\mathbf{v}, \mathbf{w}) = (c\mathbf{v}, c\mathbf{w}).$$

The linear dimension is thus equal to the sum of the dimensions of  $V$  and  $W$ . Every linear vector space of dimension greater than one can be represented by a direct sum of nonintersecting subspaces.

### 20.2.2 Multilinear Functions

Let  $V_1$ ,  $V_2$  and  $W$  be vector spaces. A function

$$f : V_1 \times V_2 \rightarrow W$$

is called **bilinear** if it is linear in each variable, i.e., if,

$$\begin{aligned} f(av_1 + bv'_1, v_2) &= af(v_1, v_2) + bf(v'_1, v_2), \\ f(v_1, av_2 + bv'_2) &= af(v_1, v_2) + bf(v_1, v'_2). \end{aligned}$$

The extension of this definition to functions of more than two variables is simple. Indeed, functions such as

$$f : V_1 \times V_2 \times \cdots \times V_n \rightarrow W \quad (20.4)$$

are called **multilinear functions**, more specifically  **$n$ -linear functions**, for which the defining relation is

$$\begin{aligned} f(v_1, \dots, av_i + bv'_i, \dots, v_n) &= af(v_1, \dots, v_i, \dots, v_n) \\ &\quad + bf(v_1, \dots, v'_i, \dots, v_n). \end{aligned}$$

An  $n$ -linear function can be multiplied by a scalar and two  $n$ -linear functions can be added; in each case the result is an  $n$ -linear function. Thus, the set of  $n$ -linear functions given in (20.4) forms a vector space denoted by  $\mathcal{L}(V_1 \times \cdots \times V_n, W)$ .

### 20.2.3 Tensor Product

Suppose that  $\tau^1 \in V_1^*$  and  $\tau^2 \in V_2^*$ , i.e.,  $\tau^1$  and  $\tau^2$  are linear real-valued functions on  $V_1$  and  $V_2$ , respectively. We can then form a bilinear real-valued function such as

$$\tau^1 \otimes \tau^2 : V_1 \times V_2 \rightarrow \mathbf{R},$$

which is represented by

$$\tau^1 \otimes \tau^2(v_1, v_2) = \tau^1(v_1)\tau^2(v_2). \quad (20.5)$$

Note that the right-hand side of (20.5) is just the product of two real numbers:  $\tau^1(v_1)$  and  $\tau^2(v_2)$ . The bilinear function  $\tau^1 \otimes \tau^2$  is called the **tensor product** of  $\tau^1$  and  $\tau^2$ . Clearly, since  $\tau^1$  and  $\tau^2$  are separately linear, so is  $\tau^1 \otimes \tau^2$ . Hence, the set of the tensor product  $\tau^1 \otimes \tau^2$  forms a vector space  $\mathcal{L}(V_1 \times V_2, \mathbf{R})$ .

Recall that the vectors  $v \in V$  can be regarded as linear functions acting on  $V^*$ . In this context, we can also construct tensor products of two vectors. For example, let  $v_1 \in V_1$  and  $v_2 \in V_2$  and define the tensor product

$$v_1 \otimes v_2 : V_1^* \times V_2^* \rightarrow \mathbf{R}$$

by

$$\mathbf{v}_1 \otimes \mathbf{v}_2(\tau_1, \tau_2) = \mathbf{v}_1(\tau_1)\mathbf{v}_2(\tau_2) = \tau_1(\mathbf{v}_1)\tau_2(\mathbf{v}_2). \quad (20.6)$$

This shows that the tensor product  $\mathbf{v}_1 \otimes \mathbf{v}_2$  can be considered a bilinear function acting on  $V_1^* \times V_2^*$ , similar to  $\tau_1 \otimes \tau_2$  being a bilinear function on  $V_1 \times V_2$ , which indicates that the set of  $\mathbf{v}_1 \otimes \mathbf{v}_2$  form a vector space  $\mathcal{L}(V_1^* \times V_2^*, \mathbf{R})$ .

Furthermore, given a vector space  $V$ , we can construct mixed types of tensor products such as

$$\mathbf{v} \otimes \tau : V^* \times V \rightarrow \mathbf{R}$$

given by

$$\mathbf{v} \otimes \tau(\phi, \mathbf{u}) = \mathbf{v}(\phi)\tau(\mathbf{u}) = \phi(\mathbf{v})\mathbf{u}(\tau), \quad (20.7)$$

where  $\mathbf{u}, \mathbf{v} \in V$  and  $\phi, \tau \in V^*$ . In a straightforward extrapolation, it is possible to develop tensor products of more than two linear functions or vectors such as

$$\mathbf{v}_1 \otimes \mathbf{v}_2 \otimes \cdots \mathbf{v}_r \otimes \tau^1 \otimes \tau^2 \otimes \cdots \tau^s, \quad (20.8)$$

which act on the vector space

$$V^* \times V^* \times \cdots \times V^* \times V \times V \times \cdots \times V,$$

where  $V^*$  appears  $r$  times and  $V$   $s$  times. Similar to the previous cases, the set of tensor products (20.8) forms a vector space denoted by

$$\mathcal{L}[(V^*)^r \times V^s, \mathbf{R}],$$

where  $(V^*)^r$  and  $V^s$  are direct products of  $V^*$  with  $r$  factors and those of  $V$  with  $s$  factors, respectively.

#### 20.2.4 General Definition of Tensors

We finally arrive at the following generalized definition of a tensor.

##### ♠ Tensor:

Let  $V$  be a vector space with a dual space  $V^*$ . Then a **tensor of type**  $(r, s)$ , denoted by  $\mathbf{T}_s^r$ , is a multilinear function

$$\mathbf{T}_s^r : (V^*)^r \times (V)^s \rightarrow \mathbf{R}.$$

The number  $r$  is called the **contravariant degree** of the tensor, and  $s$  is called the **covariant degree** of the tensor.

♠ **Tensor space:**

The set of all tensors  $\mathbf{T}_s^r$  for fixed  $r$  and  $s$  forms a vector space, called a **tensor space**, denoted by

$$\mathcal{T}_s^r(V) \equiv \mathcal{L}[(V^*)^r \times V^s, \mathbf{R}].$$

As an example, let  $\mathbf{v}_1, \dots, \mathbf{v}_r \in V$  and  $\tau^1, \dots, \tau^s \in V^*$  and define the tensor product (i.e., multilinear function)

$$\mathbf{T}_s^r \equiv \mathbf{v}_1 \otimes \dots \otimes \mathbf{v}_r \otimes \tau^1 \otimes \dots \otimes \tau^s, \quad (20.9)$$

which yields for  $\theta^1, \dots, \theta^r \in V^*$  and  $\mathbf{u}_1, \dots, \mathbf{u}_s \in V$ ,

$$\begin{aligned} & \mathbf{v}_1 \otimes \dots \otimes \mathbf{v}_r \otimes \tau^1 \otimes \dots \otimes \tau^s (\theta^1, \dots, \theta^r, \mathbf{u}_1, \dots, \mathbf{u}_s) \quad (20.10) \\ &= \mathbf{v}_1(\theta^1) \dots \mathbf{v}_r(\theta^r) \tau^1(\mathbf{u}_1) \dots \tau^s(\mathbf{u}_s) \\ &= \prod_{i=1}^r \prod_{j=1}^s \mathbf{v}_i(\theta^i) \tau^j(\mathbf{u}_j). \end{aligned}$$

Observe that each  $\mathbf{v}$  in the tensor product (20.10) requires an element  $\theta \in V^*$  to produce a real number  $\theta$ , which is why the number of factors of  $V^*$  in the direct product (20.9) equals the number of  $\mathbf{v}$ 's in the tensor product (20.10).

In particular, a tensor of type  $(0, 0)$  is defined as a **scalar**, so  $\mathcal{T}_0^0(V) = \mathbf{R}$ ; a tensor of type  $(1, 0)$ , an ordinary vector, is called a **contravariant vector**; and one of type  $(0, 1)$ , a linear function, is called a **covariant vector**. More generally, a tensor of type  $(r, 0)$  is called a **contravariant tensor** of rank (or degree)  $r$  and one of type  $(0, s)$  is called a **covariant tensor** of rank (or degree)  $s$ .

*Remark.* We can form a tensor product of two tensors  $\mathbf{T}_s^r$  and  $\mathbf{U}_\ell^k$  such as

$$\mathbf{T}_s^r \otimes \mathbf{U}_\ell^k : (V^*)^{r+k} \times V^{s+\ell} \rightarrow \mathbf{R},$$

which is a natural generalization of tensor products given in (20.5), (20.6), and (20.7). It is easy to prove that the tensor product is associative and distributive over tensor addition, but not commutative.

## 20.3 Components of Tensors

### 20.3.1 Basis of a Tensor Space

In physical applications of tensor calculus, it is necessary to choose a basis for the vector space  $V$  and one for its dual space  $V^*$  to represent the tensors by a

set of real numbers (i.e., components). The need for this process is analogous to the cases of elementary vector calculus, in which linear operators are often represented by arrays of numbers, i.e., by matrices referring to a chosen basis of the space. A basis of our tensor space  $\mathcal{T}_s^r(V) \equiv \mathcal{L}[(V^*)^r \times V^s, \mathbf{R}]$  is defined as follows.

♠ **Basis for the tensor space:**

Let  $\{e_i\}$  and  $\{\varepsilon^j\}$  be a basis in  $V$  and  $V^*$ , respectively. Then, a **basis for the tensor space**  $\mathcal{T}_s^r(V)$  is a set of all tensor products:

$$e_{i_1} \otimes \cdots \otimes e_{i_r} \otimes \varepsilon^{j_1} \otimes \cdots \otimes \varepsilon^{j_s}. \quad (20.11)$$

♠ **Components of a tensor:**

The components of any tensor  $A \in \mathcal{T}_s^r(V)$  are the real numbers given by

$$A_{j_1 \cdots j_s}^{i_1 \cdots i_r} = A(\varepsilon^{i_1}, \dots, \varepsilon^{i_r}, e_{j_1}, \dots, e_{j_s}).$$

*Remark. 1.* A useful result of the theorem is the relation

$$A = A_{j_1 \cdots j_s}^{i_1 \cdots i_r} e_{i_1} \otimes \cdots \otimes e_{i_r} \otimes \varepsilon^{j_1} \otimes \cdots \otimes \varepsilon^{j_s}.$$

2. Note that for every factor in the basis of  $\mathcal{T}_s^r(V)$ , there are  $N$  possibilities. (For instance, we have  $N$  choices for  $e_{i_1}$  in which  $i_1 = 1, 2, \dots, N$ .) Thus, the number of possible tensor products represented by (20.11) is  $N^{r+s}$ .

*Examples 1.* A tensor space  $\mathcal{T}_0^1(V)$  has a basis  $\{e_i\}$  so that an element (i.e., a contravariant vector)  $v \in \mathcal{T}_0^1(V)$  can be expanded by

$$v = v^i e_i,$$

where the real numbers  $v^i = v(\varepsilon^i)$  are called the components of  $v : V \rightarrow \mathbf{R}$ .

2. A tensor space  $\mathcal{T}_1^0(V)$  has a basis  $\{\varepsilon^j\}$  so that an element (i.e., a covariant vector)  $\tau \in \mathcal{T}_1^0(V)$  can be expanded by

$$\tau = \tau_j \varepsilon^j,$$

where the real numbers  $\tau_j = \tau(e_j)$  are called the components of  $\tau : V^* \rightarrow \mathbf{R}$ .

3. A tensor space  $\mathcal{T}_1^2(V)$  has a basis  $\{e_i \otimes e_j \otimes \varepsilon^k\}$  so that for any  $A \in \mathcal{T}_1^2(V)$  we have

$$A = A_k^{ij} e_i \otimes e_j \otimes \varepsilon^k,$$

where the real numbers

$$A_k^{ij} = \mathbf{A}(\varepsilon^i, \varepsilon^j, \mathbf{e}_k)$$

are the components of  $\mathbf{A} : V \times V \times V^* \rightarrow \mathbf{R}$ .

### 20.3.2 Transformation Laws of Tensors

The components of a tensor depend on the basis in which they are described. If the basis is changed, the components change. The relation between components of a tensor in different bases is called the **transformation law** for that particular tensor. Let us investigate this concept.

Assume two different bases of  $V$ , denoted by  $\{\mathbf{e}_i\}$  and  $\{\mathbf{e}'_i\}$ . Similarly, we denote by  $\{\varepsilon^j\}$  and  $\{\varepsilon'^j\}$  two different bases of  $V^*$ . We can find appropriate transformation matrices  $[R_i^j]$  and  $[S_k^\ell]$  that satisfy

$$\mathbf{e}'_i = R_i^j \mathbf{e}_j \quad \text{and} \quad \varepsilon'^k = S_k^\ell \varepsilon^\ell.$$

Then, for a tensor  $\mathbf{T}$  of type  $(1, 2)$ , we have

$$\begin{aligned} T_{jk}^i &= \mathbf{T}(\varepsilon'^i, \mathbf{e}'_j, \mathbf{e}'_k) = \mathbf{T}(S_\ell^i \varepsilon^\ell, R_j^m \mathbf{e}_m, R_k^n \mathbf{e}_n) \\ &= S_\ell^i R_j^m R_k^n \mathbf{T}(\varepsilon^\ell, \mathbf{e}_m, \mathbf{e}_n) \\ &= S_\ell^i R_j^m R_k^n T_{mn}^\ell, \end{aligned} \tag{20.12}$$

which is the transformation law of the components of the tensor  $\mathbf{T}$  of type  $(1, 2)$ .

Remember that in the coordinate-*dependent* treatment of tensors (as shown in Chaps. 18 and 19), the result (20.12) was considered to be the defining relation for a tensor of type  $(1, 2)$ . In other words, a tensor of type  $(1, 2)$  was defined as a collection of numbers  $T_{np}^m$  that transform to another collection of numbers  $T_{jk}^i$  according to the rule in (20.12) when the basis is changed. In our current (i.e., coordinate-*free*) treatment of tensors, it is not necessary to introduce a basis to define a tensor; a basis must be introduced only when the components of a tensor are needed. The advantage of this approach is obvious, since a  $(1, 2)$ -type tensor has 27 components in three dimensions and 64 components in four dimensions, and all of these can be represented by the single symbol  $\mathbf{T}$ .

**Remark.** Note that the above arguments do not downplay the role of components. In fact, when it comes to actual calculations, we are forced to choose a basis and manipulate the components.

### 20.3.3 Natural Isomorphism

We comment below on an important property that is specific to components of tensors  $\mathbf{A} \in \mathcal{T}_1^1(V)$ . We know that tensors  $\mathbf{A} \in \mathcal{T}_1^1(V)$  are bilinear functions such as

$$\mathbf{A} : V^* \times V \rightarrow \mathbf{R}$$

and that their components  $A_j^i$  are defined by

$$A_j^i = \mathbf{A}(\varepsilon^i, \mathbf{e}_j), \quad (20.13)$$

where each  $\varepsilon^i$  and  $\mathbf{e}_j$  is a basis of  $V^*$  and  $V$ , respectively. Now we consider the matrix

$$[A_j^i] = \begin{bmatrix} A_1^1 & A_2^1 & \cdots & A_n^1 \\ A_1^2 & & \cdots & \\ \vdots & & & \vdots \\ A_1^n & \cdots & & A_n^n \end{bmatrix}, \quad (20.14)$$

whose elements  $A_j^i$  are the same as those given in (20.13). We shall see that (20.14) is the matrix representation of a linear operator  $\mathbf{A}$  in terms of the basis  $\{\mathbf{e}_i\}$  of  $V$ , which associates a vector  $\mathbf{v} \in V$  with another  $\mathbf{u} \in V$ , i.e.,

$$\mathbf{A} : V \rightarrow V.$$

A formal statement on concerning this point is given below.

♠ **Natural isomorphism:**

For any vector space  $V$ , there is a one-to-one correspondence (called a **natural isomorphism**) between a tensor  $\mathbf{A} \in \mathcal{T}_1^1(V)$  and a linear operator  $\mathbf{A} \in \mathcal{L}(V, V)$ .

*Proof* We write the tensor  $\mathbf{A} \in \mathcal{T}_1^1(V)$  as

$$\mathbf{A} = A_j^i \mathbf{e}_i \otimes \varepsilon^j.$$

Given any  $\mathbf{v} \in V$ , we obtain

$$\begin{aligned} \mathbf{A}(\mathbf{v}) &= (A_j^i \mathbf{e}_i \otimes \varepsilon^j)(\mathbf{v}) = A_j^i \mathbf{e}_i [\varepsilon^j(\mathbf{v})] \\ &= A_j^i \mathbf{e}_i [\varepsilon^j(v^k \mathbf{e}_k)] = A_j^i \mathbf{e}_i (v^k \delta_k^j) \\ &= A_j^i v^j \mathbf{e}_i. \end{aligned} \quad (20.15)$$

Observe that the  $A_j^i v^j$  in the last term are real numbers and that  $\mathbf{e}_i \in V$ . Hence, the object  $\mathbf{A}(\mathbf{v})$  is a linear combination of bases  $\mathbf{e}_i$  for  $V$ , i.e.,

$$\mathbf{A}(\mathbf{v}) \in V.$$

Denoting  $\mathbf{A}(\mathbf{v})$  in (20.15) by  $\mathbf{u} = u^i \mathbf{e}_i$ , we find that

$$u^i = A_j^i v^j, \quad (20.16)$$

in which  $u^i, v^j$  are contravariant components of the vectors  $\mathbf{u}, \mathbf{v} \in V$ , respectively, in terms of the basis  $\{\mathbf{e}_i\}$ . The result (20.16) is identified with a matrix equation:

$$\begin{bmatrix} u^1 \\ u^2 \\ \dots \\ u^n \end{bmatrix} = \begin{bmatrix} A_1^1 & A_2^1 & \dots & A_n^1 \\ A_1^2 & & & \dots \\ \dots & & & \\ A_1^n & A_2^n & \dots & A_n^n \end{bmatrix} \begin{bmatrix} v^1 \\ v^2 \\ \dots \\ v^n \end{bmatrix}. \quad (20.17)$$

Thus we can see that given  $\mathbf{A} \in \mathcal{T}_1^1(V)$ , its components form the matrix representation  $[A_j^i]$  of a linear operator  $\mathbf{A}$  that transforms a vector  $\mathbf{v} \in V$  into another vector  $\mathbf{u} \in V$ .

Conversely, for any given linear operator on  $V$  with a matrix representation  $[A_j^i]$  in terms of a basis of  $V$ , there exists a tensor  $\mathbf{A} \in \mathcal{T}_1^1(V)$ . This suggests a one-to-one correspondence between the tensor space  $\mathcal{T}_1^1(V)$  and the vector space  $\mathcal{L}(V, V)$  comprising the linear mapping  $f: V \rightarrow V$ . ♣

A parallel discussion serves for a linear operator on  $V^*$ . In fact, for any  $\tau \in V^*$ , we have

$$\begin{aligned} \mathbf{A}(\tau) &= (A_j^i \mathbf{e}_i \otimes \varepsilon^j)(\tau) = A_j^i [\mathbf{e}_i(\tau)] \varepsilon^j \\ &= A_j^i [\mathbf{e}_i(\tau_k \varepsilon^k)] \varepsilon^j = A_j^i (\tau_k \delta_i^k) \varepsilon^j \\ &= A_j^i \tau_i \varepsilon^j, \end{aligned}$$

which means that  $\mathbf{A}(\tau)$  is a linear combination of bases  $\varepsilon^j$  for  $V^*$ , i.e.,

$$\mathbf{A}(\tau) \in V^*.$$

Denoting  $\mathbf{A}(\tau)$  by  $\theta = \theta_j \varepsilon^j$ , we obtain

$$\theta_j = A_j^i \tau_i, \quad (20.18)$$

where  $\theta_j$  and  $\tau_i$  are (covariant) components of the vectors  $\theta, \tau \in V^*$  in terms of the basis  $\{\varepsilon^i\}$ . Using the same matrix representation of  $[A_j^i]$  as in (20.17), we can rewrite (20.18) as

$$[\theta_1, \dots, \theta_n] = [\tau_1, \dots, \tau_n] \begin{bmatrix} A_1^1 & A_2^1 & \dots & A_n^1 \\ A_1^2 & & & \dots \\ \dots & & & \\ A_1^n & A_2^n & \dots & A_n^n \end{bmatrix},$$

which describes a linear mapping from a vector  $\tau \in V^*$  to another vector  $\theta \in V^*$  through the linear operator with the matrix representation  $[A_j^i]$ .

We thus conclude that there is a natural isomorphism among the three vector spaces:

$$\mathcal{T}_1^1(V) = \mathcal{L}(V^* \times V, \mathbf{R}), \quad \mathcal{L}(V, V), \quad \text{and} \quad \mathcal{L}(V^*, V^*).$$

Owing to this isomorphism, these three vector spaces can be treated as though they are the same.

#### 20.3.4 Inner Product in Tensor Language

As noted at the beginning of this chapter, our current discussion is applicable to any kind of vector space regardless of whether or not it is endowed with inner product properties. If the spaces we are dealing are inner product spaces, then all the results of Chaps. 18 and 19 are reproduced, owing the assumption that only a Euclidean space (i.e., a real inner product space; see Sect. 19.2.3) is considered there. In this subsection, we shall see that this is true, but we first have to introduce the concept of inner product in connection with our current vector spaces. Such a discussion enables us to make the correspondence between the two views of tensors clear:

tensors are *sets of index-quantities* (in Chaps. 18 and 19),

and

tensors are *linear mappings* (in Chap. 20).

Below is the definition of the inner product in the language of tensor calculus.

##### ♠ Inner product:

An inner product, denoted by  $(\ , \ )$ , is a real-valued function such as

$$(\ , \ ) : V \times V \rightarrow \mathbf{R},$$

which has the following properties:

(i) it is **nondegenerate**, i.e.,

$$(\mathbf{u}, \mathbf{v}) = 0 \quad \text{for all } \mathbf{v} \iff \mathbf{u} \equiv \mathbf{0},$$

(ii) it is **symmetric**, i.e.,  $(\mathbf{u}, \mathbf{v}) = (\mathbf{v}, \mathbf{u})$ ,

(iii) it is **positive definite**, i.e.,  $(\mathbf{u}, \mathbf{v}) > 0$  whenever  $\mathbf{u} \neq \mathbf{0}$ , and

(iv) it is **bilinear**, i.e.,  $(a\mathbf{u} + b\mathbf{v}, \mathbf{w}) = a(\mathbf{u}, \mathbf{v}) + b(\mathbf{v}, \mathbf{w})$ .

**Remark.** The set of four axioms above is a restatement of those presented in Sect. 4.1.3 for real vector spaces.

By definition, the inner product of  $\mathbf{v}$  and  $\mathbf{w}$  reads

$$(\mathbf{v}, \mathbf{w}) = (v^i \mathbf{e}_i, w^j \mathbf{e}_j) = v^i w^j (\mathbf{e}_i, \mathbf{e}_j),$$

where  $(\mathbf{e}_i, \mathbf{e}_j)$  as well as  $(\mathbf{v}, \mathbf{w})$  are certain real numbers. Then, if we establish a matrix  $[g_{ij}]$  with the entities

$$g_{ij} \equiv (\mathbf{e}_i, \mathbf{e}_j), \quad (20.19)$$

we have

$$(\mathbf{v}, \mathbf{w}) = g_{ij} v^i w^j,$$

which reproduces the previous notation (18.27) obtained via the coordinate-dependent treatment of tensors.

*Remark.* The notation in (20.19) seems to imply that the function  $(\ , \ )$  is written in terms of the dual basis  $\varepsilon^i \in V^*$  by

$$(\ , \ ) = g_{ij} \varepsilon^i \otimes \varepsilon^j. \quad (20.20)$$

However, this is not the case because (20.20) does not satisfy the symmetric property required by (ii) in the above definition of an inner product.

### 20.3.5 Index Lowering and Raising in Tensor Language

We demonstrate below another important consequence of the notation in (20.19). Since the inner product  $(\mathbf{v}, \mathbf{w})$  is a bilinear function with variables  $\mathbf{v}$  and  $\mathbf{w}$ , it is a linear function of  $\mathbf{w}$  if we fix  $\mathbf{v}$ . Assume a function  $\nu : V \rightarrow \mathbf{R}$  defined by

$$\nu(\mathbf{w}) \equiv (\mathbf{v}, \mathbf{w}).$$

Clearly,  $\nu$  is a linear function of  $\mathbf{w}$  and  $\nu \in V^*$ . Hence,  $\nu$  can be expanded by the dual basis  $\varepsilon^j \in V^*$ , which results in

$$\begin{aligned} \nu &= \nu_j \varepsilon^j = \nu(\mathbf{e}_j) \varepsilon^j \\ &= (\mathbf{v}, \mathbf{e}_j) \varepsilon^j = (v^i \mathbf{e}_i, \mathbf{e}_j) \varepsilon^j \\ &= v^i g_{ij} \varepsilon^j. \end{aligned}$$

This indicates that the components  $\nu_j$  of the linear function  $\nu$  are given by

$$\nu_j = g_{ij} v^i.$$

Denoting  $\nu_j$  by  $v_j$ , we obtain

$$v_j = g_{ij} v^i, \quad (20.21)$$

which is identified with the index lowering of  $v^i$  by the use of  $g_{ij}$ . Emphasis is placed on the fact that the result (20.21) gives a one-to-one relation between  $\mathbf{v} \in V$  and  $\nu \in V^*$  via the entities  $g_{ij} = (\mathbf{e}_i, \mathbf{e}_j)$ . That is, going from a vector  $\mathbf{v} \in V$  to its unique image  $\nu \in V^*$  is achieved by simply lowering the index of the contravariant component of  $\mathbf{v}$  through relation (20.21).

The counterpart of (20.21), index raising, is obtained by noting the fact that by hypothesis the matrix  $[g_{ij}]$  is nondegenerate. This implies the existence of the inverse matrix denoted by  $[g^{ij}]$ . Multiplying the elements  $g^{kj}$  by both sides of (20.21) yields

$$g^{kj}v_j = g^{kj}g_{ij}v^i = g^{kj}g_{ji}v^i = \delta_i^k v^i = v^k,$$

i.e.,

$$v^i = g^{ij}v_j.$$

We have thus shown that the introduction of the matrix  $[g_{ij}]$  composed of the real numbers  $g_{ij}$  defined by (20.19) provides a bridge between the two apparently different viewpoints (those in Chaps. 18, 19 and in Chap. 20) regarding tensors.



## Part VII

---

## Appendixes



# A

---

## Proof of the Bolzano–Weierstrass Theorem

### A.1 Limit Points

In this appendix we prove the **Bolzano–Weierstrass theorem**, first introduced in Sect. 2.2.2, which guarantees the existence of a **limit point** in some sets of real numbers. For a better understanding, we begin with a brief review of the basic properties of limit points.

Below is we repeat the definition of a limit point from Sect. 1.1.5.

♠ **Limit point:**

A point  $x \in \mathbf{R}$  is called a **limit point** of a set  $S \subseteq \mathbf{R}$  if every neighborhood  $V$  of  $x$  contains an element different from  $x$ .

We denote by  $\hat{S}$  the set of limit points of  $S$ . A point in  $S$  that is not a limit point of  $S$  is called an **isolated point** of  $S$ . A limit point is often referred to as a **cluster point** or **accumulation point**.

Observe that  $x \in \hat{S}$  if and only if every neighborhood of  $x$  contains an infinite number of points of  $S$ . This is so because if a neighborhood

$$V = (x - \delta, x + \delta)$$

of a limit point  $x$  contains only a finite number of points, say  $a_1, a_2, \dots, a_n$ , where  $a_i \neq x$ , then there is a positive number  $\varepsilon$  such that

$$\varepsilon = \min_{1 \leq i \leq n} |a_i - x|.$$

Since  $x$  is a limit point of  $S$ , there is a point  $a \in S$  such that  $a \neq x$  and  $|x - a| < \varepsilon$ . This means that  $a \in V$  but  $a \neq a_i$  for any  $i$ , which contradicts the assumption that  $V$  contains only  $n$  points of  $S$ . The implication in the other direction is obvious.

A **finite set** cannot have a limit point, since any neighborhood of a limit point must contain an infinite number of elements of the set. On the other hand, an **infinite set** may or may not have a limit point.

## A.2 Cantor Theorem

The previous discussion raises the question: When does a set possess a limit point? The following theorem serves as a lemma to answer this question. Meanwhile we denote by  $\ell(I) = b - a$  the length of any closed interval  $I = [a, b]$  with  $a \leq b$ .

♠ **Cantor theorem:**

Let  $(I_n)$  be a sequence of nonempty, closed, and bounded intervals. If  $I_{n+1} \subseteq I_n$  for every  $n \in \mathbf{N}$ , then the intersection  $\bigcap_{n=1}^{\infty} I_n$  is not empty. Furthermore, if

$$\inf\{\ell(I_n) : n \in \mathbf{N}\} = 0,$$

then  $\bigcap_{n=1}^{\infty} I_n$  is a single point.

*Proof* Suppose  $I_n = [a_n, b_n]$  and  $I = \bigcap_{n=1}^{\infty} I_n$ . Using the nested property of the intervals  $I_n$ , we have

$$m \geq p \Rightarrow I_m \subseteq I_p \Rightarrow a_p \leq a_m < b_m \leq b_p. \quad (\text{A.1})$$

Clearly, the set  $S = \{a_n : n \in \mathbf{N}\} \subset \mathbf{R}$  is not empty and is bounded above by  $b_1$ . Hence, the set  $S$  has a least upper bound, which we call  $x$ . If we can prove that  $x \in I$ , we can conclude that  $I$  is not empty. In fact, this can be proven by observing that

$$x \in I_k \quad \text{for all } k \in \mathbf{N},$$

i.e.,

$$a_k \leq x \leq b_k \quad \text{for all } k \in \mathbf{N}. \quad (\text{A.2})$$

First, it is obvious from the definition of  $x$  that  $a_k \leq x$  for all  $k$ . Second,  $b_k$  for arbitrary  $k$  satisfies  $a_n \leq b_k$  for all  $n \in \mathbf{N}$ , i.e.,  $b_k$  is an upper bound of  $S$ . In fact, if  $n \leq k$  then, by (A.1),  $a_n \leq a_k \leq b_k$ ; and if  $n > k$  then, again by (A.1),  $a_n \leq b_n \leq b_k$ . Finally, it follows that  $x \leq b_k$  for all  $k \in \mathbf{N}$ , since  $x$  is a *least* upper bound of  $S$ , whereas  $b_k$  is just an upper bound of  $S$ . Thus we can conclude that (A.2) is true.

Now we consider the second statement in the above theorem. Suppose that

$$\inf\{\ell(I_n) : n \in \mathbf{N}\} = 0,$$

and let  $x, y \in I$ . It then follows that  $x, y \in I_n$  for every  $n$ , which implies that

$$|x - y| \leq \ell(I_n) \quad \text{for all } n \in \mathbf{N},$$

so

$$|x - y| \leq \inf\{\ell(I_n) : n \in \mathbf{N}\} = 0.$$

This means that  $x = y$ , i.e., the interval

$$I = \bigcap_{n=1}^{\infty} I_n$$

is a single point. ♣

### A.3 Bolzano–Weierstrass Theorem

We are now ready to prove the Bolzano–Weierstrass theorem, which gives us sufficient conditions for the existence of a limit point in a set.

♠ **Bolzano–Weierstrass theorem:**

Every infinite and bounded subset of  $\mathbf{R}$  has at least one limit point in  $\mathbf{R}$ .

*Proof* Let  $S$  be an infinite and bounded set of real numbers. Being bounded,  $S$  is contained in some bounded closed interval  $I_0 = [a_0, b_0]$ . First we bisect  $I_0$  into the two subintervals

$$I'_0 = \left[ a_0, \frac{a_0 + b_0}{2} \right], \quad I''_0 = \left[ \frac{a_0 + b_0}{2}, b_0 \right].$$

Since  $S \subseteq (I'_0 \cup I''_0)$  is infinite, at least one of the two sets  $S \cap I'_0$  and  $S \cap I''_0$  is infinite. Let  $I_1 = [a_1, b_1] = I'_0$  if  $S \cap I'_0$  is infinite; otherwise let  $I_1 = I''_0$ . We then have

$$I_1 \subseteq I_0, \quad \ell(I_1) = \frac{b_0 - a_0}{2}.$$

Now we bisect  $I_1$  into the subintervals

$$I'_1 = \left[ a_1, \frac{a_1 + b_1}{2} \right], \quad I''_1 = \left[ \frac{a_1 + b_1}{2}, b_1 \right],$$

one of which necessarily intersects  $S$  in an infinite set. Let  $I_2 = [a_2, b_2] = I'_1$  if  $S \cap I'_1$  is infinite; otherwise let  $I_2 = I''_1$ . Continuing in this fashion, we obtain the intervals  $I_i = [a_i, b_i]$ ,  $0 \leq i \leq n$ , which satisfy

$$I_i \subseteq I_{i-1}, \quad \ell(I_i) = \frac{b_0 - a_0}{2^i},$$

and the fact that  $S \cap I_i$  is infinite. We bisect  $I_n$  again to obtain

$$I'_n = \left[ a_n, \frac{a_n + b_n}{2} \right], \quad I''_n = \left[ \frac{a_n + b_n}{2}, b_n \right].$$

Since  $I_n = I'_n \cup I''_n$  and  $S \cap I_n$  is infinite, either  $S \cap I'_n$  is infinite wherein we

set  $I_{n+1} = [a_{n+1}, b_{n+1}] = I'_n$ , or  $S \cap I''_n$  is infinite where  $I_{n+1} = I''_n$  is chosen. Now we see that

$$I_{n+1} \subseteq I_n, \quad \ell(I_{n+1}) = \frac{b_0 - a_0}{2^{n+1}} = \frac{\ell(I_n)}{2}$$

and that  $S \cap I_{n+1}$  is infinite. By induction, therefore, it is show that there exists a sequence  $(I_n)$  of nonempty, closed, and bounded intervals. In view of Cantor's theorem, we see that the intersection  $\bigcap_{n=1}^{\infty} I_n$  contains at least one single point  $x$ . We now complete our proof by showing that  $x \in \hat{S}$ .

Suppose that  $x \in \bigcap_{n=1}^{\infty} I_n$ . Given any  $\varepsilon > 0$ , we can choose  $n \in \mathbf{N}$  so that

$$\frac{b_0 - a_0}{2^n} < \varepsilon,$$

or equivalently,

$$|I_n| < \varepsilon.$$

This, together with the fact that  $x \in I_n$  for all  $n$ , implies that

$$I_n \subset (x - \varepsilon, x + \varepsilon).$$

Since  $I_n$  contains an infinite number of elements of  $S$ , so does the neighborhood  $(x - \varepsilon, x + \varepsilon)$  of  $x$ ; hence  $x \in \hat{S}$ . ♣

## B

---

### Dirac $\delta$ Function

#### B.1 Basic Properties

In this appendix, we review the properties and various expressions of **Dirac's  $\delta$  function**. The first thing to be noted is that the  $\delta$  function is not a function at all. A function is a rule that assigns another number to each number in a set of numbers. However, the  $\delta$  function as used in physics is rather a shorthand notation for a complicated limiting process whose use greatly simplifies calculations. It takes on a meaning only when it appears under an integral sign, in which case it behaves as

$$\int_{-\infty}^{\infty} f(x)\delta(x)dx = f(0), \quad (\text{B.1})$$

For the special case of  $f(x) \equiv 1$ . We have

$$\int_{-\infty}^{\infty} \delta(x)dx = 1. \quad (\text{B.2})$$

If the singular point is located at an arbitrary point  $x_0$ , then

$$\int_{-\infty}^{\infty} f(x)\delta(x - x_0)dx = f(x_0). \quad (\text{B.3})$$

Except at the singular point  $x_0 = 0$ ,

$$\delta(x) = 0. \quad (\text{B.4})$$

Thus  $\delta(x)$  vanishes at all points where its argument is not zero, but at that one point it is undefined. Nevertheless its behavior near this point is all that matters.

Let  $\delta_a(x)$  be a set of functions parametrized by the index  $a$  that has the properties

$$\begin{aligned}\lim_{a \rightarrow 0} \delta_a(x) &= 0 \quad \text{for all } x \neq 0, \\ \lim_{a \rightarrow 0} \int_{-\infty}^{+\infty} f(x) \delta_a(x) dx &= f(0).\end{aligned}\tag{B.5}$$

In precise terms the original equations defining the  $\delta$  function must be interpreted as standing for the limiting processes of (B.5).

## B.2 Representation as a Limit of Function

In what follows, we look at several sets of functions that are endowed the properties described in (B.5).

### 1. The limit of a box function

The simplest example is the function  $\delta_c(x)$  defined (for  $c > 0$ ) by

$$\delta_c(x) \equiv \begin{cases} 1/c & \text{for } |x| \leq c/2, \\ 0 & \text{for } |x| > c/2. \end{cases}\tag{B.6}$$

Clearly,  $\lim_{c \rightarrow 0} \delta_c(x) = 0$  at all  $x \neq 0$  and  $\int_{-\infty}^{+\infty} \delta_c(x) dx = 1$  independent of  $c$ . In addition, we have

$$\lim_{c \rightarrow 0} \int_{-\infty}^{\infty} f(x) \delta_c(x) dx = f(0),\tag{B.7}$$

which can be shown formally for continuous functions  $f(x)$  as

$$\begin{aligned}\lim_{c \rightarrow 0} \int_{-\infty}^{\infty} f(x) \delta_c(x) dx &= \lim_{c \rightarrow 0} \int_{-c/2}^{c/2} f(x) \delta_c(x) dx = \lim_{c \rightarrow 0} \frac{1}{c} \int_{-c/2}^{c/2} f(x) dx \\ &= \lim_{c \rightarrow 0} \frac{f(\xi c)}{c} \int_{-c/2}^{c/2} dx = \lim_{c \rightarrow 0} f(\xi c).\end{aligned}$$

In the last line, the mean value theorem of integral calculus was employed with the definition  $-1/2 < \xi < 1/2$ . Letting  $c \rightarrow 0$ , we obtain (B.7).

### 2. The limit of a Gaussian function

The sequence of the Gaussian distribution function

$$\delta_a(x) \equiv \frac{1}{a\sqrt{\pi}} e^{-x^2/a^2}$$

provides another representation of the  $\delta$  function. Note that  $\lim_{a \rightarrow 0} \delta_a(x) = 0$  at all  $x \neq 0$ ,  $\int_{-\infty}^{+\infty} \delta_a(x) dx = 1$  independent of  $a$ , and  $\lim_{a \rightarrow 0} \int_{-\infty}^{\infty} f(x) \delta_a(x) dx = f(0)$ . The entire contribution to the integral, as  $a \rightarrow 0$ , comes from the neighborhood of  $x = 0$ . Therefore, we may write symbolically,

$$\delta(x) = \lim_{a \rightarrow 0} \delta_a(x) = \lim_{a \rightarrow 0} \frac{1}{a\sqrt{\pi}} e^{-x^2/a^2}.$$

### 3. The limit of a Lorentzian function

Another useful representation for the  $\delta$  function is

$$\delta(x) = \lim_{\epsilon \rightarrow 0} \delta_{\epsilon}(x) \equiv \lim_{\epsilon \rightarrow 0} \frac{1}{\pi} \frac{\epsilon}{x^2 + \epsilon^2},$$

which the reader can work out as in the above example.

### 4. The $n \rightarrow \infty$ limit of a sequence of functions

The final representation of the  $\delta$  function is slightly different from the preceding three and plays a central role in the proof of the Weierstrass theorem, as demonstrated in Appendix C. It is defined as

$$\delta_n(x) = \begin{cases} c_n(1 - x^2)^n & \text{for } 0 \leq |x| \leq 1 \quad (n = 1, 2, 3 \dots), \\ 0 & \text{for } |x| > 1, \end{cases} \quad (\text{B.8})$$

where the constant  $c_n$  must be determined so that

$$\int_{-1}^1 \delta_n(x) dx = 1. \quad (\text{B.9})$$

The functions  $\delta_n(x)$  form a sequence whose limit is a  $\delta$  function. This representation of the  $\delta$  function differs from the others in that the defining parameter  $n$  increases to infinity, rather than decreasing continuously to zero.

## B.3 Remarks on Representation 4

We show below that representation 4 above has the conditions (B.5) required for identification with Dirac's  $\delta$  function. At first, we determine the normalization constant  $c_n$ . From the hypothesis (B.9), we have

$$\frac{1}{c_n} = \int_{-1}^1 (1 - x^2)^n dx = 2 \int_0^1 (1 - x^2)^n dx. \quad (\text{B.10})$$

Making the change of variable  $x = \sin \theta$ , we obtain

$$\frac{1}{c_n} = 2 \int_{\pi/2}^0 \cos^{2n+1} \theta d\theta = \frac{2^{n+1} n!}{1 \cdot 3 \cdot 5 \cdots (2n+1)}, \quad (\text{B.11})$$

which becomes

$$c_n = \frac{(2n+1)!}{2^{2n+1} (n!)^2}. \quad (\text{B.12})$$

Next we consider the asymptotic behavior of  $c_n$  as  $n \rightarrow \infty$ . It follows from (B.10) that

$$\frac{1}{c_n} = 2 \int_0^1 (1-x^2)^n dx \geq 2 \int_0^{1/\sqrt{n}} (1-x^2)^n dx, \quad (\text{B.13})$$

since  $1/\sqrt{n}$  for all  $n = 1, 2, \dots$  and the integrand is positive throughout  $[0, 1]$ . Now we consider the function

$$g(x) \equiv (1-x^2)^n - (1-nx^2).$$

Since  $g(0) = 0$  and

$$g'(x) = 2nx [1 - (1-x^2)^{n-1}] > 0 \quad \text{for all } 0 < x \leq 1,$$

$g(x)$  must be monotonically increasing in the interval  $[0, 1]$ . Therefore,  $g(x) \geq 0$ , or equivalently,

$$(1-x^2)^n \geq 1-nx^2$$

for all  $x$  in  $[0, 1]$ . Using this inequality in (B.13), we have

$$\frac{1}{c_n} \geq 2 \int_0^{1/\sqrt{n}} (1-nx^2)^n dx = \frac{4}{3\sqrt{n}} > \frac{1}{\sqrt{n}},$$

i.e.,

$$c_n < \sqrt{n}. \quad (\text{B.14})$$

This result implies that the limit  $n \rightarrow \infty$  of the function  $\delta_n(x)$  given in (B.8) equals zero for all  $x \neq 0$ .

Finally, we examine the validity of the relation:

$$\lim_{n \rightarrow \infty} \int_{-\infty}^{\infty} f(x) \delta_n(x) dx = f(0). \quad (\text{B.15})$$

To prove this, it suffices to verify that the contribution to the integral  $\int_{-1}^1 \delta_n(x) dx$  comes increasingly from the neighborhood surrounding the origin as  $n \rightarrow \infty$ . Note that for  $0 < \varepsilon < 1$ ,

$$\int_{-1}^{-\varepsilon} \delta_n(x) dx = \int_{\varepsilon}^1 \delta_n(x) dx, \quad (\text{B.16})$$

since  $\delta_n(x)$  is an even function of  $x$ . Now

$$\int_{\varepsilon}^1 \delta_n(x) dx < \sqrt{n} \int_{\varepsilon}^1 (1-x^2)^n dx < \sqrt{n}(1-\varepsilon^2)^n(1-\varepsilon) < \sqrt{n}(1-\varepsilon^2)^n, \quad (\text{B.17})$$

where we employ the fact that  $(1-x^2)^n$  in the interval  $[\varepsilon, 1]$  takes its maximum at  $x = \varepsilon$ . It is obvious that the decreasing behavior of the term  $(1-\varepsilon^2)^n$  with  $n$  dominates increasing behavior of the term  $\sqrt{n}$ , so that

$$\lim_{n \rightarrow \infty} \int_{\varepsilon}^1 \delta_n(x) dx = 0. \quad (\text{B.18})$$

The results (B.16) and (B.18) justify the desired relation (B.15).



## C

---

### Proof of Weierstrass Approximation Theorem

♠ **Weierstrass approximation theorem:**

If a function  $f(x)$  is continuous on the closed interval  $[a, b]$ , there exists a polynomial

$$G_n(x) = \sum_{k=0}^n c_k x^k \quad (\text{C.1})$$

that converges uniformly to  $f(x)$  on  $[a, b]$ .

To prove this, we may assume without loss of generality that  $f(x)$  is defined on  $[0, 1]$  and that  $f(0) = f(1) = 0$ . Outside the interval  $[0, 1]$ , we may define  $f(x)$  to be identically zero. Then, the relevant polynomial (C.1) is given by the integral form as

$$G_n(x) = \int_{-1}^1 f(x+t) \delta_n(t) dt, \quad 0 \leq x \leq 1. \quad (\text{C.2})$$

Here  $\delta_n(t)$  is the sequence of the functions represented by

$$\delta_n(t) = \begin{cases} c_n(1-t^2)^n & \text{for } -1 \leq t \leq 1, \\ 0 & \text{for } |t| > 1, \end{cases}$$

where  $c_n$  is

$$c_n = \frac{(2n+1)!}{2^{2n+1}(n!)^2}, \quad \text{so that } \int_{-1}^1 \delta_n(t) dt = 1.$$

(In fact, the sequence  $\delta_n(t)$  as  $n \rightarrow \infty$  does have the properties characterizing a  **$\delta$  function**; see Appendix B). Since  $f(x)$  is assumed to vanish outside the interval  $[0, 1]$ , (C.2) can be rewritten as

$$G_n(x) = \int_{-x}^{1-x} f(x+t) \delta_n(t) dt.$$

By a change of variable  $t \rightarrow t - x$ , we obtain

$$G_n(x) = \int_0^1 f(t) \delta_n(t-x) dt = \int_0^1 f(t) c_n [1 - (t-x)^2]^n dt.$$

This last integral shows that  $G_n(x)$  is a polynomial of degree  $2n$  in  $x$ . In what follows, we prove that the sequence of polynomials given by  $\{G_1(x), G_2(x), \dots\}$  converges uniformly to  $f(x)$ .

Since  $f(x)$  is continuous on  $[0, 1]$ , there exists an appropriate infinitesimal  $\delta$  such that for a given  $\varepsilon > 0$ ,

$$|f(x + \delta) - f(x)| < \varepsilon$$

for all  $x$  in  $[0, 1]$ . Now, we use (C.2) for  $G_n(x)$  to obtain the quantity

$$\begin{aligned} |G_n(x) - f(x)| &= \left| \int_{-1}^1 [f(x+t) - f(x)] \delta_n(t) dt \right| \\ &\leq \int_{-1}^1 |f(x+t) - f(x)| \delta_n(t) dt, \end{aligned} \quad (\text{C.3})$$

where  $\delta_n(t) \geq 0$  on  $t \in [0, 1]$ . If the last term in (C.3) vanishes as  $n \rightarrow \infty$ , the proof of the theorem is complete.

To show this, we break up the range of integration into three parts,

$$\int_{-1}^1 = \int_{-1}^{-\gamma} + \int_{-\gamma}^{\gamma} + \int_{\gamma}^1,$$

where  $\gamma$  is a certain infinitesimal number. Since  $f(x)$  is continuous on a closed interval, it is bounded there. Let the maximum value of  $|f(x)| = M$ . Then the last term of integrals becomes

$$\begin{aligned} \int_{\gamma}^1 |f(x+t) - f(x)| \delta_n(t) dt &\leq \int_{\gamma}^1 |f(x+t)| \delta_n(t) dt + \int_{\gamma}^1 |f(x)| \delta_n(t) dt \\ &\leq 2M \int_{\gamma}^1 \delta_n(t) dt < 2M\sqrt{n}(1 - \gamma^2)^n, \end{aligned} \quad (\text{C.4})$$

where we have used the inequality (B.17). Similar arguments yield

$$\int_{-1}^{-\gamma} |f(x+t) - f(x)| \delta_n(t) dt < 2M\sqrt{n}(1 - \gamma^2)^n. \quad (\text{C.5})$$

Finally, the remaining integral,  $\int_{-\gamma}^{\gamma}$ , is estimated by using the continuity of  $f(x)$ , which guarantees that for any  $\varepsilon$  we can find an infinitesimal  $\gamma$  that satisfies

$$|t| < \gamma \Rightarrow |f(x+t) - f(x)| < \varepsilon.$$

This yields

$$\int_{-\gamma}^{\gamma} |f(x+t) - f(x)| \delta_n(t) dt < \varepsilon \int_{-\gamma}^{\gamma} \delta_n(t) dt < \varepsilon, \quad (\text{C.6})$$

since  $\int_{-\gamma}^{\gamma} \delta_n(t) dt < 1$ .

Collecting the results of (C.3)–(C.6), we have

$$|G_n(x) - f(x)| < 4M\sqrt{n}(1 - \gamma^2)^n + \varepsilon.$$

The value of  $\sqrt{n}(1 - \gamma^2)^n$  for  $0 < \gamma < 1$  can be set arbitrarily small for large enough  $n$  and, in particular, smaller than  $\varepsilon$ . Therefore, there exists an  $N$  such that

$$n > N \Rightarrow |G_n(x) - f(x)| < \varepsilon$$

for any arbitrarily small preassigned  $\varepsilon$  where  $N$  does not depend on  $x$ . This means that the sequence of polynomials  $G_n(x)$  converges uniformly to the continuous function  $f(x)$  on  $[0, 1]$ , which completes the proof. We emphasize that the above discussion holds for an arbitrary continuous function on an arbitrary finite closed interval  $[a, b]$ , as was indicated at the outset.

*Remark.* It should be emphasized that our initial hypothesis that  $f(0) = f(1) = 0$  imposes *no* limitation on the validity of the proof. To see this, we now suppose that  $f(x)$  is defined on  $[a, b]$ . Then, the function  $g(x)$  defined by

$$g\left(\frac{x-a}{b-a}\right) \equiv f(x)$$

yields  $f(a) = g(0)$  and  $f(b) = g(1)$ , where any  $x$  in the interval  $[a, b]$  corresponds to  $z$  in  $[0, 1]$ . Furthermore, by introducing the function

$$h(z) = g(z) - g(0) - z[g(1) - g(0)]$$

for  $z$  in  $[0, 1]$ , we have  $h(0) = 0$  and  $h(1) = 0$ . We can show that the polynomial  $G_n(x)$  that approximates the original function  $f(x)$  also approximates the modified function  $h(z)$  by replacing the variable  $x$  in  $G_n(x)$  by  $z$ .



## D

---

### Tabulated List of Orthonormal Polynomial Functions

#### Hermite Polynomials $H_n(x)$

Orthogonality:

$$\int_{-\infty}^{\infty} e^{-x^2} H_m(x) H_n(x) dx = 2^n n! \sqrt{\pi} \delta_{mn}.$$

Rodrigues formula:

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n} (e^{-x^2}).$$

Differential equation:

$$\frac{d^2}{dx^2} H_n(x) - 2x \frac{d}{dx} H_n(x) + 2n H_n(x) = 0.$$

Recurrence formula:

$$H_{n+1}(x) - 2x H_n(x) + 2n H_{n-1}(x) = 0.$$

Generating functions:

$$g(t, x) = e^{2xt - t^2} = \sum_{n=0}^{\infty} \frac{H_n(x)}{n!} t^n.$$

#### Laguerre Polynomials $L_n^\nu(x)$

Orthogonality:

$$\int_0^\infty x^\nu e^{-x} L_m^\nu(x) L_n^\nu(x) dx = \frac{\Gamma(n + \nu + 1)}{\Gamma(n + 1)} \delta_{mn}.$$

Rodrigues formula:

$$L_n^\nu(x) = \frac{x^{-\nu}}{n!} e^x \frac{d^n}{dx^n} (e^{-x} x^{\nu+n}).$$

Differential equation:

$$x \frac{d^2}{dx^2} L_n^\nu(x) + (\nu + 1 - x) \frac{d}{dx} L_n^\nu(x) + n L_n^\nu(x) = 0.$$

Recurrence formula:

$$(n+1)L_{n+1}^\nu(x) - (2n+\nu+1)L_n^\nu(x) - (n+\nu)L_{n-1}^\nu(x) = 0.$$

Generating functions:

$$g(t, x) = \frac{e^{-xt/(1-t)}}{(1-t)^{\nu+1}} = \sum_{n=0}^{\infty} L_n^\nu(x) t^n.$$

### Jacobi Polynomials $G_n^{(\nu, \mu)}(x)$

Orthogonality:

$$\int_{-1}^1 (1+x)^\mu (1-x)^\nu G_m^{(\nu, \mu)}(x) G_n^{(\nu, \mu)}(x) dx = \frac{2^{\mu+\nu+1} \Gamma(n+\mu+1) \Gamma(n+\nu+1)}{n! (2n+\mu+\nu+1) \Gamma(n+\nu+\mu+1)} \delta_{mn}.$$

Rodrigues formula:

$$G_n^{(\nu, \mu)}(x) = \frac{(-1)^n}{2^n n!} (1-x)^{-\nu} (1+x)^{-\mu} \frac{d^n}{dx^n} [(1-x)^{\nu+n} (1+x)^{\mu+n}].$$

Differential equation:

$$\begin{aligned} (1-x^2) \frac{d^2}{dx^2} G_n^{(\nu, \mu)}(x) + [\mu - \nu - (\nu + \mu + 2)x] \frac{d}{dx} G_n^{(\nu, \mu)}(x) \\ + n(n + \nu + \mu + 1) G_n^{(\nu, \mu)}(x) = 0. \end{aligned}$$

Recurrence formula:

$$\begin{aligned} G_0^{(\nu, \mu)}(x) = 1, \quad G_1^{(\nu, \mu)}(x) = \frac{1}{2} \{(\nu + \mu + 2)x + (\nu^2 - \mu^2)\}, \\ 2(n+1)(n+\nu+\mu+1)(2n+\nu+\mu) G_{n+1}^{(\nu, \mu)}(x) \\ - (2n+\nu+\mu+1) [(2n+\nu+\mu)(2n+\nu+\mu+2)x + \nu^2 - \mu^2] G_n^{(\nu, \mu)}(x) \\ - 2(n+\nu)(n+\mu)(2n+\nu+\mu+2) G_{n-1}^{(\nu, \mu)}(x) = 0. \end{aligned}$$

Generating functions:

$$\begin{aligned} g(t, x) &= \frac{2^{\nu+\mu}}{(1-2xt+t^2)^{1/2} \left\{1-t+(1-2xt+t^2)^{1/2}\right\}^{\nu} \left\{1+t+(1-2xt+t^2)^{1/2}\right\}^{\mu}} \\ &= \sum_{n=0}^{\infty} G_n^{(\nu, \mu)}(x) t^n. \end{aligned}$$

### Gegenbauer Polynomials $C_n^{\lambda}(x)$

Orthogonality:

$$\int_{-1}^1 (1-x^2)^{\lambda-\frac{1}{2}} C_m^{\lambda}(x) C_n^{\lambda}(x) dx = \frac{\sqrt{\pi} \Gamma(n+2\lambda) \Gamma(\lambda-\frac{1}{2})}{n!(n+\lambda) \Gamma(2\lambda) \Gamma(\lambda)} \delta_{mn}.$$

Rodrigues formula:

$$C_n^{\lambda}(x) = \frac{(-1)^n \Gamma(n+2\lambda) \Gamma[\lambda+\frac{1}{2}]}{2^n n! \Gamma[n+\lambda+\frac{1}{2}] \Gamma(2\lambda)} (1-x^2)^{-\lambda+\frac{1}{2}} \frac{d^n}{dx^n} \left[ (1-x^2)^{n+\lambda-\frac{1}{2}} \right].$$

Differential equation:

$$(1-x^2) \frac{d^2}{dx^2} C_n^{\lambda}(x) - (2\lambda+1)x \frac{d}{dx} C_n^{\lambda}(x) + n(n+2\lambda) C_n^{\lambda}(x) = 0.$$

Recurrence formula:

$$(n+1) C_{n+1}^{\lambda}(x) - 2(n+\lambda)x C_n^{\lambda}(x) - (n+2\lambda-1) C_{n-1}^{\lambda}(x) = 0.$$

Generating functions:

$$g(t, x) = \frac{1}{(1-2xt+t^2)^{\lambda}} = \sum_{n=0}^{\infty} C_n^{\lambda}(x) t^n.$$

### Legendre Polynomials $P_n(x)$

Orthogonality:

$$\int_{-1}^1 P_m(x) P_n(x) dx = \frac{2}{2n+1} \delta_{mn}.$$

Rodrigues formula:

$$P_n(x) = \frac{(-1)^n}{2^n n!} \frac{d^n}{dx^n} [(1-x^2)^n].$$

Differential equation:

$$(1-x^2)\frac{d^2}{dx^2}P_n(x) - 2x\frac{d}{dx}P_n(x) + n(n+1)P_n(x) = 0.$$

Recurrence formula:

$$(n+1)P_{n+1}(x) - (2n+1)xP_n(x) + nP_{n-1}(x) = 0.$$

Generating functions:

$$g(t, x) = \frac{1}{(1-2xt+t^2)^{1/2}} = \sum_{n=0}^{\infty} P_n(x)t^n.$$

### Chebyshev Polynomials of the First Kind $T_n(x)$

Orthogonality:

$$\int_{-1}^1 \frac{T_m(x)T_n(x)}{\sqrt{1-x^2}}dx = \frac{\pi}{2}\delta_{mn} (1 + \delta_{m0}\delta_{n0}).$$

Rodrigues formula:

$$T_n(x) = \frac{(-2)^n n! \sqrt{\pi}}{(2n)!} (1-x^2)^{\frac{1}{2}} \frac{d^n}{dx^n} \left[ (1-x^2)^{n-\frac{1}{2}} \right].$$

Differential equation:

$$(1-x^2)\frac{d^2}{dx^2}T_n(x) - x\frac{d}{dx}T_n(x) + n^2T_n(x) = 0.$$

Recurrence formula:

$$T_{n+1}(x) - 2xT_n(x) + T_{n-1}(x) = 0.$$

Generating functions:

$$g(t, x) = \frac{1-t^2}{1-2xt+t^2} = \sum_{n=1}^{\infty} 2T_n(x)t^n + T_0(x).$$

### Chebyshev Polynomials of the Second Kind $U_n(x)$

Orthogonality:

$$\int_{-1}^1 \sqrt{1-x^2}U_m(x)U_n(x)dx = \frac{\pi}{2}\delta_{mn} (1 - \delta_{m0}\delta_{n0}).$$

Rodrigues formula:

$$U_n(x) = \frac{(-2)^n(n+1)!\sqrt{\pi}}{(2n+1)!}(1-x^2)^{-\frac{1}{2}}\frac{d^n}{dx^n}\left[(1-x^2)^{n+\frac{1}{2}}\right].$$

Differential equation:

$$(1-x^2)\frac{d^2}{dx^2}U_n(x) - 3x\frac{d}{dx}U_n(x) + n(n+2)U_n(x) = 0.$$

Recurrence formula:

$$U_{n+1}(x) - 2xU_n(x) + U_{n-1}(x) = 0.$$

Generating functions:

$$g(t, x) = \frac{1}{1-2xt+t^2} = \sum_{n=0}^{\infty} U_n(x)t^n.$$



---

# Index

- Abscissa of absolute convergence, 411, 412
- Abscissa of convergence, 410–412, 415
- Absolute convergence, 32, 38, 40, 42, 67, 214, 383, 411, 412, 424
- Absolute convergence of an infinite series, 33, 38, 40
- Absolute maximum theorem, 207
- Absolute minimum theorem, 208
- Absolutely divergent series, 42
- Accumulation point, 6, 657
- Active transformation, 570, 571
- Addition, 83, 640, 641, 643, 646
- Addition formula for analytic continuation, 254
- Addition identity, 83
- Addition of complex number, 74
- Addition of tensor, 586
- Addition of vector, 74
- Addition theorem of trigonometric function, 251
- Additive inverse, 83
- Adjoint operator, 502
- Admissibility condition, 450, 459
- Admissibility constant, 450
- Airfoil, 335, 336
- Aliasing, 394
- Almost everywhere, 156, 158, 160, 168, 174, 456, 474
- Alternating sequence, 42
- Alternating series, 42
- Alternating series test, 42
- Amplitude modulation, 404
- Analytic continuation, 215, 216, 246, 247, 249, 251–254, 284, 409, 420, 421, 432
- Analytic continuations of each other, 247, 248
- Analytic function, 102, 125, 186–188, 191–194, 198, 201, 202, 204–208, 210, 213, 217, 220, 236, 237, 244, 251, 254, 259, 263, 286, 289, 290, 305, 306, 409, 432, 436, 438
- Analyticity, 188, 190, 192, 193, 195, 261, 312, 321, 412, 432
- Analyticity at infinity, 313
- Angle-preserving, 305, 306
- Angular momentum, 583, 591, 596
- Angular velocity, 596
- Anharmonic ratio, 322
- Antisymmetric part of tensor, 590
- Antisymmetric tensor, 589–591
- Approximation coefficient, 472, 478–480
- Area under the graph, 145, 382
- Associated Legendre function, 112
- Associative, 83, 387, 389, 646
- Associativity, 83
- Asymptotically stable critical point, 528
- Auto-correlation function, 390
- Autonomous system, 528
- Axial vector, 582
- Banach space, 86, 87
- Banach's fixed point theorem, 173
- Basis, 79, 87, 88, 612, 642, 646, 649, 650
- Basis of a Hilbert space, 91
- Basis of tensor space, 647, 648

- Bernoulli equation, 505, 506
- Bernoulli's theorem, 230, 232
- Bessel equation, 505
- Bessel function, 385, 403
- Bessel inequality, 79, 82, 89, 96, 104, 357
- Beta function, 108
- Bilateral Laplace transform, 433
- Bilinear, 643, 651
- Bilinear function, 644, 645, 648, 651, 652
- Bilinear transformation, 316, 317, 321, 324, 325
- Binomial theorem, 24
- Bit reversal, 398
- Bit-reversing process, 399, 401
- Blasius' formula, 230–232
- Bolzano-Weierstrass' theorem, 26, 27, 80, 657, 659
- Boundary point, 7
- Bounded above, 3, 20, 21, 36–38, 43
- Bounded almost everywhere, 160
- Bounded below, 3, 20, 21
- Bounded closed interval, 5
- Bounded open interval, 5
- Bounded real sequence, 19–21, 26, 27
- Bounded set, 3
- Branch, 226, 241–244, 252, 272, 273, 276, 421
- Branch at infinity, 313
- Branch cut, 226, 243, 244, 441, 442
- Branch line, 244
- Branch point, 235, 243, 244, 252, 282, 441
- Brownian motion, 550, 551
- Cantor set, 155, 157
- Cantor's theorem, 658
- Cardinal number, 155
- Carrier wave, 404–406
- Cartesian basis, 79
- Cartesian coordinate, 317, 319
- Cartesian coordinate system, 567, 568, 576, 580, 582, 602, 603
- Cartesian product of vector spaces, 643
- Cartesian space, 78
- Cartesian tensor, 578, 589, 596, 601
- Cartesian tensor of the first order, 576, 577
- Cartesian tensor of the fourth order, 585, 600
- Cartesian tensor of the second order, 578
- Cartesian tensor of the third order, 585
- Cartesian vector, 576–578, 582
- Cauchy criterion, 25–27, 31, 36, 38, 53, 55, 59, 62, 64, 79
- Cauchy criterion for convergence, 31
- Cauchy criterion for uniform convergence, 53
- Cauchy inequality, 208
- Cauchy principal value, 294
- Cauchy problem, 552, 553
- Cauchy sequence, 25–28, 36, 54, 55, 79, 80, 82, 86, 91–94, 170, 171, 174, 175, 496
- Cauchy's integral formula, 124, 205–207, 213, 217, 432
- Cauchy's test for improper integrals, 68, 425, 428
- Cauchy's theorem, 198–201, 205, 210, 220, 262, 274, 291
- Cauchy-Riemann relation, 189, 194, 308, 309, 332
- Causality requirement, 300
- Center, 533, 535
- Central limit theorem, 175–178, 180
- Characteristic curve of a PDE, 542
- Characteristic equation of a PDE, 542
- Characteristic function, 176, 178, 180
- Characteristic polynomial, 529
- Chebyshev polynomial of the First Kind, 674
- Chebyshev polynomial of the first kind, 119, 129–131, 133, 135
- Chebyshev polynomial of the Second Kind, 674
- Chebyshev polynomial of the second kind, 119
- Chebyshev's inequality, 157, 158
- Christoffel symbol, 621–625, 627, 628, 630–632, 635
- Christoffel symbol of the first kind, 623
- Christoffel symbol of the second kind, 623
- Circle of convergence, 214–217, 221, 222, 246, 250, 253, 254, 256
- Circulation (of fluid flow), 229

- Clairaut equation, 488
- Closed set, 7
- Closedness, 83, 430
- Closure, 5, 7
- Cluster point, 6, 657
- Cofactor, 524, 572–575
- Column-vector notation, 510
- Commutative, 83, 387, 389, 640, 646
- Complement, 2
- Complementary minor of an element of a matrix, 571
- Complementary set, 2, 7, 8, 149, 150, 156, 157
- Complementary system, 514
- Complete, 73, 79, 80, 86–89, 91, 101, 170, 173, 515
- Complete analytic function, 248
- Complete integral of an ODE, 487
- Complete orthonormal set of functions, 73, 97, 98, 109, 463, 464
- Complete orthonormal set of polynomials, 101, 103–105, 117, 119
- Complete orthonormal vectors, 89
- Completeness of wavelets, 462, 463
- Complex conjugate, 75, 533, 641
- Complex function, 185, 186
- Complex sphere, 311
- Complex vector space, 74–76, 83
- Component of a tensor, 565, 576, 578
- Conditional convergence, 32–35, 37, 42
- Conditional convergence of an improper integral, 67, 412
- Conditional convergence of an infinite series, 33
- Conformal, 305
- Conformal mapping, 306–308, 310–313, 315–317, 321, 322, 324, 328, 330, 331, 333–335
- Conjugate harmonic function, 192
- Conjugate linear, 75
- Conjugate space, 641
- Conservation law of current flow, 537
- Conservation law of momentum, 230, 231, 373
- Conservation of a functional equation, 250, 253
- Contact point, 5–9
- Continuity of complex function, 186
- Continuity theorem (for characteristic functions), 180
- Continuity theorem (of integrals), 67
- Continuous approximation, 472, 476
- Continuous function, 47, 50
- Continuous on the left, 48
- Continuous on the right, 48
- Contraction, 587–589, 591, 596, 598, 618
- Contraction mapping, 173, 174
- Contraction mapping theorem, 173–175
- Contrapositive proof, 9
- Contravariant basis vector, 603, 604, 606, 621
- Contravariant component of a tensor, 608–610, 612, 613, 618, 619, 629
- Contravariant component of a vector, 607, 611, 613, 617, 627, 628, 631, 650, 653
- Contravariant degree of tensor, 645
- Contravariant local basis vector, 603
- Contravariant tensor, 646
- Contravariant vector, 646, 647
- Convergence test for alternating series, 42
- Convergence almost everywhere, 156, 171
- Convergence of a real sequence, 17
- Convergence of a sequence of vectors, 80
- Convergence of an improper integral, 67
- Convergence of an infinite series, 30, 33
- Convergence of Laplace integral, 408–411, 422, 424–427, 430–432, 435
- Convergence test, 29, 38, 42
- Convolution, 387–389, 451, 453, 459, 479
- Convolution integral, 447, 476
- Convolution theorem, 387
- Coordinate, 88, 567
- Coordinate axis, 567
- Coordinate transformation, 566, 567, 570, 577, 580
- Corner, 50
- Correlation function, 388
- Countable set, 154, 155
- Covariant basis vector, 603–605, 608, 613, 617, 619, 621

- Covariant component of a tensor, 609, 610, 613, 619, 632
- Covariant component of a vector, 607, 608, 611, 613, 617, 627, 628, 631, 635, 650
- Covariant constant, 633
- Covariant degree of tensor, 645
- Covariant derivative, 628–633
- Covariant differentiation, 634, 635
- Covariant local basis vector, 602, 603
- Covariant tensor, 646
- Covariant vector, 608, 646, 647
- Critical point of an autonomous system, 527–534
- Critical point of conformal mapping, 308, 310
- Cross ratio, 322, 325
- Cross-correlation function, 388–390
- Curvature tensor, 635
- Curvilinear coordinate system, 565, 601–603, 605, 607, 611, 615, 617, 621
- Cut, 244
- Cylindrical coordinate system, 612, 616, 621, 624
  
- D'Alembertian, 552
- Damped harmonic oscillator, 383
- Damping time constant, 446
- Darboux's inequality, 196, 209, 211
- Decomposition algorithm, 478, 479
- Decreasing sequence, 20
- Derivative (of a complex function), 186, 187
- Derivative (of a real function), 48
- Determinant of a matrix, 571
- Difference, 2
- Differentiability (of a complex function), 186, 188
- Differentiability (of a real function), 48
- Diffusion constant, 551
- Diffusion equation, 371, 545, 550, 551, 561, 562
- Diffusion operator, 546, 550
- Dilatation parameter, 451, 454
- Dilation equation, 468
- Dimension of a vector space, 88
- Dirac's  $\delta$ -function, 661–663, 667
- Direct product (of vector spaces), 643, 645, 646
- Direct product (of vectors), 578
- Direct proof, 9
- Direct sum of vector spaces, 643
- Directed cosine, 568
- Direction field, 490, 494, 526
- Dirichlet boundary condition, 332–334, 540, 556
- Dirichlet problem for the diffusion equation, 551
- Dirichlet problems for the Laplace equation, 548
- Dirichlet theorem, 360
- Dirichlet's conditions for the Fourier series convergence, 341, 347
- Dirichlet's function, 149, 155, 156, 172
- Dirichlet's integral, 353, 358
- Dirichlet's kernel, 354
- Dirichlet's theorem, 341
- Discrete Fourier transform, 391–394, 396, 398, 400, 401
- Discrete wavelet, 460, 462, 463
- Discrete wavelet transform, 460–462, 467, 472, 476, 478
- Disjoint interval, 141, 146
- Disjoint set, 2, 144, 145
- Dispersion relation, 297–302
- Displacement vector, 600
- Distance, 84, 174, 639
- Distance function, 84, 85
- Distribution, 176–180
- Distributive, 83, 387, 389, 646
- Divergence (as a vector operation), 631
- Divergent sequence, 18, 19
- Divergent series, 32, 33, 35, 42
- Divergent test, 32
- Dominated convergence theorem, 158, 160, 161, 165, 166
- Dual basis, 642, 652
- Dual space, 641, 642, 645, 646
- Dummy index, 566, 626, 628, 629
- Dyadic grid arrangement, 461
- Dyadic grid wavelet, 461
  
- Eigenenergy, 136
- Eigenfrequency, 557
- Eigenfunction, 136, 501, 504, 505, 557
- Eigenvalue, 501, 504, 505, 530–534, 544

- Eigenvector, 530–534
- Einstein tensor, 636
- Einstein's field equation, 635–637
- Elasticity theory, 600
- Elastisity theory, 585
- Electric conductivity, 598
- Electromagnetic field, 599
- Element, 1–5, 7, 74, 76, 83
- Elliptic class of PDEs, 544
- Elliptic coordinate, 319
- Elliptic coordinate system, 565
- Elliptic integral of the first kind, 326
- Empty set, 1, 141, 142
- Entire function, 191, 209, 313
- Enumerable, 154
- Equal, 2
- Equality almost everywhere, 156, 158, 174, 175, 456, 474
- Equivalent, 10
- Essential singularity, 233, 235–240, 282
- Essential singularity at infinity, 313
- Euclidean space, 3, 74, 75, 515, 614, 639, 640, 651
- Euler's formula, 108
- Euler-Fourier formula, 340, 344
- Existence theorem, 491, 495, 498, 515
- Expected value of a random variable, 143, 176
- Explicit solution of an ODE, 484
- Exponential order, 423–427, 431
- Extended definition of conformal mappings, 312
- Extended real number, 3
- False, 9
- Fast Fourier transform, 396, 397
- Fast Fourier transform (FFT), 396, 398, 399, 401
- Fast orthogonal wave transform, 478
- Fast wavelet transform, 460, 477–480
- Father wavelet, 463, 470, 477
- Fejér's integral, 353, 358
- Fejér's theorem, 355, 360, 361
- Finite set, 1, 140, 154, 155, 657
- First shifting theorem, 415
- First-order Cartesian pseudotensor, 582, 585
- First-order linear homogeneous ODE, 484
- First-order linear homogeneous PDE, 541
- Fixed point in  $L^p$ , 174
- Flat Riemann space, 614
- Flat space, 634–636
- Fluid flow, 229
- Focus, 533
- Four potential, 599
- Four-current density, 600
- Four-vector, 76
- Four-velocity, 636
- Fourier coefficient, 95–98, 105
- Fourier cosine series, 344, 345, 350
- Fourier integral, 383
- Fourier integral representation, 378
- Fourier integral theorem, 379, 380
- Fourier series, 95, 96, 339–341, 360, 363, 366, 377
- Fourier sine series, 344, 350
- Fourier transform, 378, 382, 390, 391, 406, 435, 559
- Fourier transform in three dimension, 384
- Fourier transform in two dimension, 385
- Fractional transformation, 316
- Fraunhofer diffraction, 401, 403
- Frequency modulation, 404
- Fresnel cosine integral, 279, 281
- Fresnel sine integral, 279, 281
- Fubini's theorem, 162–164, 166, 167, 173, 175, 176, 178, 180
- Fubini-Hobson-Tonelli theorem, 164
- Function element, 247
- Function of exponential order, 423–427, 431
- Function space, 172, 173
- Fundamental matrix, 518, 519, 521
- Fundamental mixed tensor, 611
- Fundamental sequence, 25
- Fundamental system of solutions, 516–523
- Fundamental tensor, 585
- G.l.b., 4
- Gamma function, 108, 254
- Gauss notation, 106
- Gegenbauer polynomial, 119, 673
- General analytic function, 248
- General relativity theory, 634, 636

- General solution of a differential equation, 316, 372, 375, 487–489, 522, 530, 531, 533, 534, 540, 542, 545, 553, 554, 557, 559
- Generalized Fourier coefficient, 91, 95
- Generalized Fourier series, 95
- Generating function, 113, 114, 124–126, 470
- Generating function of Chebyshev polynomials of the first kind, 674
- Generating function of Chebyshev polynomials of the second kind, 675
- Generating function of Gegenbauer polynomials, 673
- Generating function of Hermite polynomials, 125, 671
- Generating function of Jacobi polynomials, 673
- Generating function of Laguerre polynomials, 125, 672
- Generating function of Legendre polynomials, 113, 114, 125, 674
- Generating function of the multiresolution analysis, 470
- Geometric curvature, 634
- Gibbs phenomenon, 347, 365, 366
- Goursat's formula, 206–208, 261, 265
- Gradient, 631
- Gradient of a scalar, 631
- Gradient of a vector, 580
- Gram-Schmidt orthogonalization method, 103, 105, 114, 505
- Greatest lower bound, 4, 411
- Green's function, 558, 559
- Gutzmer's theorem, 227
- Haar discrete wavelet, 462, 472, 473
- Haar wavelet, 450, 458, 467, 473
- Half-range Fourier series, 344, 347
- Harmonic function, 191, 195, 546, 549, 550
- Harmonic series, 35, 36, 39
- Heat flow, 550
- Heat flux, 561
- Hermite equation, 502
- Hermite polynomial, 117, 125, 127, 135, 671
- Hermitian operator, 503
- Hilbert space, 73, 74, 79–83, 87–92, 95, 98, 352
- Hilbert space theory, 352
- Hilbert transforms pair, 295–298
- Holomorphic, 188
- Hooke's law, 600
- Hyperbolic class of PDEs, 544, 546, 552, 553
- Hyperharmonic series, 36
- Identically distributed, 176, 177, 179
- Identity vector, 74
- If and only if, 10
- Imaginary part of a complex function, 185
- Implicit solution of an ODE, 485, 486, 490
- Improper integral, 66–68, 302, 412, 420, 422, 425–428, 430, 433
- Improper node, 530, 531
- Improper rotation, 580–585
- Incomplete inner product space, 81
- Incompressible, 228
- Increasing sequence, 19
- Independent random variable, 176, 177
- Index lowering, 652
- Index raising, 652
- Inertia tensor, 596–598
- Infimum, 4, 8, 140
- Infinite series, 29–33, 37, 38, 40, 42, 96, 105, 109, 221, 339
- Infinite series of functions, 62–64, 227, 281, 340, 342, 362, 496
- Infinite set, 1, 154, 155, 157, 657
- Initial value problem, 419, 491–493, 495, 497–499, 510, 513, 527, 552, 554, 555
- Inner measure, 150, 157
- Inner product, 73, 75, 76, 78, 80, 87, 89, 96, 97, 352, 588, 639, 641, 651
- Inner product (in tensor calculus), 651, 652
- Inner product notation, 502, 503
- Inner product space, 75–82, 86, 87, 640, 651
- Integral curve, 488–490
- Integral equation, 492
- Integral function, 313
- Integral of PDE, 540

- Interior point, 7
- Intersection, 2, 87, 247, 658, 660
- Interval, 4
- Invariant, 609
- Invariant tensor, 585
- Inverse Fourier transform, 378, 379, 384, 387, 395, 396, 406, 471
- Inverse Fourier transformation, 435
- Inverse Laplace transform, 408, 409, 432, 434, 436, 439–441, 444, 446, 448, 558
- Inverse matrix, 523, 574, 620, 653
- Inverse of discrete Fourier transform, 392, 393
- Inverse of the two-sided Laplace transform, 434, 435
- Inverse wavelet transform, 456–458, 460
- Inversion (as a bilinear transformation), 321, 327–329
- Inversion (as an improper rotation), 581, 582
- Irrotational, 228, 231
- Isolated point, 6–8, 95, 149, 212, 657
- Isolated singularity, 233–236, 239, 252, 262, 263, 313
- Isomorphism, 98, 649
- Isomorphism between  $\ell^2$  and  $L^2$ , 98, 99
- Isotropic tensor, 585, 600
  
- Jacobi matrix, 122
- Jacobi polynomial, 118, 672
- Jacobian determinant, 309, 388, 555, 617
- Jordan's lemma, 270, 438, 441
- Joukowski airfoil, 336
- Joukowski transformation, 335, 336
  
- Kinetic energy, 597
- Kramers-Kronig relations, 299
- Kronecker's delta, 78, 610
- Kutta-Joukowski's theorem, 228–231
  
- L'Hôpital's rule, 12, 13, 239, 280, 282, 370, 417, 423
- L.u.b., 3
- Laguerre polynomial, 117, 118, 125, 671
- Lamé constants, 600
- Langevin's function, 283
- Laplace transform, 408
- Laplace equation, 191, 192, 228, 331–334, 545, 546, 548–550
- Laplace integral, 408, 410, 411, 421, 422, 428, 429, 432, 433
- Laplace operator, 546
- Laplace transform, 407–409, 412, 414–418, 420, 422, 432, 444–447, 558, 559
- Laplace transform of derivative, 419
- Laplace transform of integral, 420
- Laplacian, 545, 546, 548–550, 559, 630, 631
- Laurent series expansion, 219–224, 226, 233–235, 238, 239, 260, 265, 266
- Least upper bound, 3, 658
- Lebesgue convergence theorem, 173, 175, 178, 180, 181
- Lebesgue integrable, 152, 153, 161, 173, 174
- Lebesgue integral, 139, 141, 144, 147, 149, 151–155, 158, 161, 162, 167, 172, 175, 176
- Lebesgue measurable function, 167, 170
- Lebesgue measurable set, 157
- Lebesgue measure, 149–151, 154–156, 176
- Lebesgue sum, 151–153
- Left-hand limit, 46, 48, 408
- Left-handed coordinate system, 568, 582
- Legendre polynomial, 105–107, 109, 112–114, 119, 125, 127, 136, 137, 673
- Legendre's equation, 501
- Levi-Civita symbol, 584, 588
- Lift, 336
- Lift force, 228
- Limit, 18
- Limit cycle, 536, 538
- Limit inferior, 21, 22, 140
- Limit of a function, 45
- Limit point, 6, 18, 657
- Limit superior, 21, 22, 40–42, 140
- Limit test for convergence, 38, 39, 43
- Limit test for divergence, 39, 40, 43
- Line element, 490
- Linear autonomous system, 528
- Linear differential equations, 407
- Linear function, 640

- $n$ -linear function, 644
- Linear homogeneous ODE, 484, 500
- Linear homogeneous PDE, 541
- Linear homogeneous system of ODEs, 514, 516, 517, 524, 528
- Linear independence, 76, 79, 82, 88, 103, 375, 515–517, 519, 520, 522, 523, 569, 642
- Linear inhomogeneous ODE, 505, 506
- Linear inhomogeneous PDE, 541
- Linear inhomogeneous system of ODEs, 514, 516
- Linear mapping (of vector spaces), 640, 650, 651
- Linear ODE, 484
- Linear partial differential equation (PDE), 540
- Linear space, 640
- Linear transformation, 315, 316, 514, 543, 544
- Liouville's formula, 518, 521, 524
- Liouville's theorem, 209, 238, 313
- Lipschitz condition, 174, 495, 497, 500, 512, 515, 526
- Lipschitz constant, 495, 512
- Local basis vectors, 602
- Localization theorem, 368
- Logarithmic residue, 286, 287
- Logistic equation, 506
- Lower bound, 3
- Lower limit, 21
- Lower Riemann-Darboux integral, 140
  
- Möbius transformation, 316, 322
- Magnetic susceptibility, 598
- Maximum, 4
- Maxwell equation, 599, 637
- Maxwell-Boltzmann distribution, 177
- Mean convergence, 95, 97, 105, 351–353, 355–357, 360, 361
- Mean value of a random variable, 143, 144, 176
- Mean value theorem, 58, 204, 560, 562, 662
- Measurable set, 151, 157, 160, 164, 165, 167
- Measure, 141
- $\alpha$ -measure, 141, 144, 146–148
- Message wave, 404–406
  
- Method of inversion, 327
- Method of variation of constant parameters, 522
- Metric coefficient, 616
- Metric space, 84, 85
- Metric tensor, 611–614, 617–619, 621, 623, 624, 626, 630, 631, 633–635, 637
- Metric vector space, 84
- Mexican hat wavelet, 450–452, 455, 467
- Minimax property, 129
- Minimum, 4
- Minkowski's inequality, 93, 169–171, 175
- Minor of elements of a matrix, 571
- Mixed component of a tensor, 609–612, 618
- Modified summation convention, 603, 605, 606
- Moment of inertia, 597, 609
- Monotone convergence theorem, 158–161, 165, 166, 171
- Monotonic sequence, 20
- Monotonically decreasing sequence, 20
- Monotonically increasing sequence, 19
- Morera's theorem, 210
- Mother wavelet, 467–470, 477
- Multilinear function, 644–646
- Multiplication of complex number, 74
- Multiply connected region, 200
- Multipole, 136
- Multiresolution algorithm, 478
- Multiresolution analysis, 463, 464, 466–470, 474, 477
- Multiresolution analysis equation, 468
- Multiresolution representation, 472
- Multivalued function, 226, 240–244, 252, 318, 409, 420, 441
  
- Natural boundary, 249, 253, 256
- Natural isomorphism, 649–651
- Natural pairing, 643
- Necessary and sufficient condition, 10
- Necessary condition, 9
- Neighborhood, 5–8, 10, 12, 18, 24, 66, 67, 187, 188, 218, 233–236, 239, 244, 250, 260, 286, 308, 312, 313, 316, 369, 528, 657, 660

- Neumann boundary condition, 332, 540, 556
- Neutrally stable critical point, 528
- Newtonian field of gravity, 636
- Noise reduction, 457, 458
- Non-degenerate, 651
- Non-isolated singularity, 235
- Nonhomogeneous linear partial differential equation, 541
- Nonlinear differential equation, 484, 505, 506, 522, 538, 637
- Nonoverlapping sets, 141
- Norm, 76, 80, 85–87, 89, 91, 94–96, 174, 352, 473, 510, 639
- $p$ -norm, 85–87, 168, 175
- Normal distribution, 177–180
- Normed space, 85–87
- Null measure, 155, 157
- Nyquist critical frequency, 393
  
- Once-subtracted dispersion relation, 300
- One-sided derivative, 49
- One-sided limit, 46
- Open set, 7
- Order of differential equation, 483
- Order of zero of function, 233
- Ordinary differential equation (ODE), 174, 483
- Orthogonal basis, 88
- Orthogonal complement, 465
- Orthogonal curvilinear coordinate, 317
- Orthogonal decomposition, 466
- Orthogonal polynomial, 114–117, 119, 121–124, 126, 128, 129
- Orthogonal relation, 579
- Orthogonality, 73, 78, 79, 82, 88–90, 103, 105, 107, 127
- Orthogonality relation, 115, 119, 120, 129, 133
- Orthonormal basis, 73, 78, 88, 463, 464, 466–471
- Orthonormality, 78, 101
- Orthonormality of wavelets, 462, 463
- Outer measure, 150, 156
- Outer product, 578, 588, 595, 610
  
- Parabolic class of PDEs, 544
- Parallelogram law, 78, 87
  
- Parseval's identity, 90, 97, 98, 104, 302, 356, 357, 362, 383, 390
- Parseval's identity (for wavelet transform), 457, 460, 474
- Partial differential equation (PDE), 371, 539
- Partial sum, 30, 31, 35–38, 43, 64, 99, 104, 109, 256, 345, 346, 353, 358, 363, 365, 366, 368, 369, 496, 500
- Particular solution of differential equation, 487, 506, 522, 523, 553
- Partition, 140, 152
- Passive transformation, 570, 571
- Path independence, 198
- Permutation symbol, 584
- Phase space, 527, 538
- Picard's method, 491, 492, 497, 499
- Piecewise continuous function, 48, 50, 362–364, 385, 417
- Piecewise smooth function, 50, 360, 363, 364, 379, 380, 386
- Plancherel's identity, 390
- Point, 1
- Point at infinity, 237, 238, 244, 312, 313, 320, 322, 329
- Point of equilibrium of an autonomous system, 527
- Pointwise convergence, 51, 52, 54, 60, 62, 95, 158, 173, 360, 363, 364
- Pointwise limit, 51
- Poisson's equation, 547, 636
- Poisson's integral formula, 213
- Polar coordinate system, 108, 194, 244, 280, 310, 315, 317–319, 323, 403, 549, 565
- Polar vector, 582
- Pole, 233–238, 240
- Pole at infinity, 313
- Positive definite, 651
- Potential field, 114, 136, 137, 228, 229, 232, 333–335, 583, 599, 636
- Power spectrum, 383, 390, 405, 406
- Pre-Hilbert space, 86, 87
- Primitive integral of an ODE, 487
- Principal part in the Laurent series, 222, 234, 235, 238
- Principal value integral, 68, 206, 293–296, 300
- Probability, 143, 176, 177

- Probability density, 136, 176, 177  
Probability density function, 143  
Probability distribution function, 143, 144  
Product of inertia, 597  
Proof by contradiction, 9  
Proper node, 532, 533  
Proper rotation, 581, 582, 584, 585  
Proper subset, 2  
Pseudotensor, 580, 582  
Pseudovector, 582, 583, 590  
Pyramid algorithm, 478  
Pythagorean formula, 73
- Quantum mechanics theory, 135  
Quotient law, 592
- Radius of convergence, 102, 214–219, 223, 245, 246, 252, 256, 257  
Random variable, 143, 144, 176, 177, 179, 180  
Range convention, 566, 568  
Ratio method, 263, 266, 267, 278  
Ratio test for convergence, 40, 44  
Rational function, 132, 237, 238, 267, 268, 271, 273  
Real part of a complex function, 185  
Real vector space, 83  
Reality condition, 298  
Rearrangement, 34, 35, 38  
Reconstruction algorithm, 478–480  
Rectangular Cartesian coordinate system, 565, 567, 570, 606  
Recurrence formula for orthogonal polynomials, 119–121, 125–127, 129, 671–675  
Recurrence relation (for analytic continuation), 254  
Recurrence relation (of gamma functions), 255  
Recurrence relation (of scaling functions), 468  
Reduced system, 514  
Refinement equation, 468  
Reflection, 581–583  
Region of analyticity, 206–208, 250  
Region of convergence of the Laplace integral, 410, 425, 427, 430–435, 444
- Region of the existence, 249  
Regular, 188  
Regular analytic, 188  
Regular part in the Laurent series, 222  
Regular point, 527  
Removable singularity, 233, 234, 236  
Residue, 259, 260, 263–268, 272–274, 281, 282, 285, 286  
Residue theorem, 259, 260, 263, 267, 269, 274, 277–279, 281, 286, 437  
Riccati equation, 506  
Ricci curvature, 637  
Ricci scalar, 636  
Ricci tensor, 636  
Ricci's theorem, 630, 633, 634  
Riemann integrable, 140  
Riemann integrable function, 158, 172  
Riemann integral, 139, 140, 144, 152, 153, 158, 172, 175  
Riemann space, 614  
Riemann sphere, 311–313  
Riemann sum, 144, 153  
Riemann surface, 241–243, 421, 422, 441  
Riemann tensor, 635, 636  
Riemann's theorem, 35  
Riemann-Darboux integral, 140  
Riemann-Lebesgue theorem, 358, 364, 369  
Riemann-Stieltjes integral, 144  
Riemann-zeta function, 284  
Riesz-Fisher's theorem, 96, 98  
Right-hand limit, 46, 48, 408  
Right-handed coordinate system, 567  
Rigid rotation of coordinate axes, 567–570, 575, 576, 580, 583, 585  
Rodrigues formula, 106, 107, 114–119, 121, 123, 124, 127, 129, 671–675  
Root test for convergence, 41, 44, 253  
Rotation (as a vector operation), 632  
Rotation (of fluid flow), 229  
Rotation with reflection, 581  
Rouché's theorem, 210, 290–292
- Saddle point, 531, 532  
Sampling theorem, 394  
Scalar, 83, 579, 609, 646  
Scalar curvature, 636, 637

- Scalar multiplication, 74, 83, 640, 641, 643
- Scalar product, 75, 579, 604, 607, 611, 613, 617, 618
- Scale factor, 308, 310, 321, 616, 617, 634
- Scale-dependent thresholding, 458
- Scaling function, 463, 464, 467–471, 474, 476, 479
- Scaling function coefficient, 468–470, 474, 477
- Scaling function space, 467
- Schrödinger equation, 136
- Schwarz differential equation, 316
- Schwarz Lemma, 211
- Schwarz principle of reflection, 254, 257
- Schwarz's inequality, 76, 77, 89
- Schwarz-Christoffel transformation, 325–327, 329, 332, 333
- Second shifting theorem, 416
- Second-order Cartesian pseudotensor, 584
- Second-order linear homogeneous PDE, 543
- Secular equation, 529
- Self-adjoint operator, 503
- Semiclosed interval, 5
- Sequence of partial sum, 30, 31, 43, 104, 109
- Sequence of the remainder, 30
- Set, 1
- Shuffled sequence, 28, 37
- Signal approximation, 472
- Signal detail, 472
- Simple Laplace development, 572
- Simple set, 146
- Simple statement, 9
- Simply connected region, 199
- Single-valued function, 84, 240, 241, 243, 247, 248, 288–290, 408, 414, 421, 484, 510, 614
- Singular line, 253
- Singular point of an autonomous system, 527
- Singular solution of ODE, 488, 489
- Singularity, 188, 233
- Skew-symmetric, 589
- Smooth function, 50
- Solution of ODE, 484
- Solution of PDE, 540
- $L^2$  space, 81, 87, 92, 95, 98, 99, 352
- $L^p$  space, 86, 87
- $\ell^2$  space, 80, 87, 91, 92, 95, 96, 98, 99
- $\ell^p$  space, 86, 87
- Specific heat, 562
- Spectrum of Sturm-Liouville system, 501
- Spherical coordinate system, 384, 549, 550, 616
- Spherical harmonic function, 109, 111, 112
- Spiral point, 533, 534
- Square-integrable function, 81, 92, 94, 95, 98, 101, 352, 357, 358
- Stability of critical point, 527, 528
- Stable critical point, 528
- Step function, 81, 144–148, 365, 366, 416, 417, 436, 440, 445
- Strain tensor, 600
- Stream function, 228
- Stress tensor, 600
- Strictly decreasing sequence, 20
- Strictly increasing sequence, 19
- Strictly stable critical point, 528
- Sturm-Liouville equation, 500–502, 505
- Sturm-Liouville operator, 500, 503, 504
- Sturm-Liouville system, 501, 504
- Subinterval, 5
- Subset, 1
- Subtraction of tensor, 586
- Successive approximation, 492, 493, 495, 499
- Sufficient condition, 9
- Sum of infinite series, 30
- Sum of infinite series of functions, 62
- $\alpha$ -summable, 146, 147
- Summation convention, 566, 568, 602
- Support, 144–146
- Supremum, 3, 4, 8, 11, 140
- Symmetric Cartesian tensor of the second order, 597
- Symmetric part of tensor, 590
- Symmetric tensor, 589
- Taylor series expansion, 49, 102, 212, 217–219, 222–225, 239, 246, 254, 260, 263, 265, 266, 284, 294, 500, 528
- Tensor, 565, 645

- Tensor of the first order, 609
- Tensor of the second order, 609
- Tensor of zero order, 579, 609
- Tensor product, 644–647
- Tensor space, 646–648, 650
- Thermal conductivity, 372, 561
- Total variation of argument, 288–290
- Trajectory, 526–528, 530
- Transfinite number, 155
- Translation parameter, 451, 453
- Tree algorithm, 478
- Triangle inequality, 77, 92, 169
- Trigonometric Fourier series, 109, 339, 340
- Trigonometric series, 339, 340, 355, 360
- True, 9
- Tunnel diode, 536
- Two-scale relation, 468, 469, 477, 478
- Two-sided Laplace integral, 433, 434
- Two-sided Laplace transform, 433–435, 444
- Unbounded open interval, 5
- Unbounded set, 3
- Uniform convergence (of complex-function sequence), 213, 214, 217, 227
- Uniform convergence (of Fourier series), 341, 342, 352, 355, 359–362, 366
- Uniform convergence (of improper integral), 67, 68, 380
- Uniform convergence (of Laplace integrals), 425–427, 429–432
- Uniform convergence (of polynomial sequence), 101, 102, 104
- Uniform convergence (of real-function sequence), 52–68, 95, 158, 172, 497
- Union, 2
- Uniqueness of the integral, 198
- Uniqueness theorem (for analytic continuation), 250, 252, 254
- Uniqueness theorem (for characteristic function), 176, 179
- Uniqueness theorem (for solution of ODE), 136, 491, 497, 498, 515, 516, 526
- Uniqueness theorem of the Dirichlet problem (for the diffusion equation), 552
- Uniqueness theorem of the Dirichlet problem (for the Laplace equation), 548
- Unit scalar, 83
- Unitary space, 75
- Universal gravitational constant, 636
- Universal set, 2
- Unstable critical point, 528
- Upper bound, 3
- Upper limit, 21
- Upper Riemann-Darboux integral, 140
- Van der Pol equation, 538
- Vanishing order, 11
- Variation of argument, 288
- Vector, 74
- Vector space, 73, 74, 83, 640
- Velocity potential, 228
- Wave equation, 373, 545, 552, 555, 556, 558, 559
- Wave operator, 546, 552, 553
- Wavelet, 449
- Wavelet analysis, 449
- Wavelet coefficient, 469, 472, 477, 478
- Wavelet space, 467
- Wavelet transform, 451–456, 458, 460
- Weierstrass approximation theorem, 101
- Weierstrass'  $M$  test, 63
- Weierstrass' test for improper integral, 68, 426
- Weight function, 75, 115, 452, 500, 502, 505
- Wiener-Kinchin's theorem, 389, 390
- Wigner-Seitz cell, 348
- Winding number, 262, 263, 291
- Wronskian, 518, 521
- Wronsky determinant, 518, 521
- Zero of function, 105, 122, 129, 130, 132, 133, 207, 209, 210, 212, 264, 286, 289, 367, 404
- Zero scalar, 83
- Zero vector, 78, 83, 89, 90
- Zeros of function, 264
- Zeta function, 36

